MASTER EN SCIENCES MATHÉMATIQUES

Comparaison de méthodes pour la détermination du nombre de classes en classification automatique

André, Paul

Award date:
1997

Facultés Universitaires Notre-Dame de la Paix
Namur
Facultés des Sciences - Département de Mathématique

Comparaison de Méthodes pour la Détermination du Nombre de Classes en Classification Automatique

Mémoire présenté pour l'obtention du grade
de Licencié en Sciences
Mathématiques
par

Paul ANDRE

Promoteur : A. Hardy

Année académique 1996-1997

Pour commencer, je tiens à remercier Monsieur le Professeur A. Hardy, promoteur de ce mémoire, pour le soutien , la disponibilité et la gentillesse dont il a fait preuve tout au long de ce travail.

Ensuite, je tiens également à remercier mes camarades, et en particulier ceux de l'option statistique, grâce à qui j'ai toujours pu travailler dans la bonne humeur.

Enfin, je remercie toute ma famille qui m'a permi de tenir le coup dans les moments difficiles, et Nathalie, sans qui je ne serais certainement pas arrivé jusqu'ici.

A tous, merci.

Résumé

Le but de la classification automatique est de décomposer un ensemble donné de n objets décrits par un ensemble de p caractéristiques, en un nombre relativement restreint de classes d'objets "semblables".
Un des problèmes fondamentaux de la classification automatique est la détermination du nombre de classes.
Dans un de leurs travaux, Milligan et Cooper ont classé trente méthodes de détermination du nombre de classes. Le but du mémoire est d'analyser les six méthodes les mieux classées à partir d'ensembles de données tests. On les compare ensuite à trois autres méthodes basées sur le critère des hypervolumes, et qui ont été développées dans l'Unité de Statistique du département de Mathématique des Facultés Universitaires de Namur.
Les résultats montrent que, sur les exemples testés, les méthodes basées sur le critère des hypervolumes donnent généralement de meilleurs résultats.

Abstract

The aim of classification is to decompose a given set of n objects described by a set of p features, in a relatively small number of clusters of similar objects. One of the fundamentals problems in classification is the determination of the number of true clusters.
In a previous work, Milligan and Cooper ranked thirty methods to determine the number of true clusters. The aim of this work is to analyse the six best ranked methods by using sets of data artificially generated to have a particular structure. Then, we compare these methods with three other methods based on the hypervolumes criterion, and that have been developed in the Statistical Unit of the Mathematics Department of the University of Namur. The results show that, for the tested examples, the methods based on the hypervolumes criterion generally give the best results.

Table des matières

Chapitre 1

La classification

1.1 Introduction

La classification automatique, encore appelée classification non-supervisée ou analyse typologique, apporte une réponse au problème suivant :

comment décomposer une population donnée d'individus
ou d'objets décrits par un ensemble de caractéristiques
en un certain nombre de groupes homogènes.

Elle permet alors de simplifier une réalité complexe par la constitution de groupes d'individus "semblables".
Cette préoccupation est commune à de nombreuses disciplines: biologie, zoologie, archéologie, botanique, géographie, géologie, agronomie, médecine, psychiatrie, psychologie, sociologie, anthropologie, linguistique, documentation automatique, intelligence artificielle, sciences de la gestion des entreprises, ... Pour exemples, les astronomes classifient les étoiles, les botanistes classifient les plantes et les linguistes classifient les langages. Un nom est alors donné à chaque classe (par exemple, en psychiatrie : la classe des maniaques, la classe des schizophrènes, ...). D'autres exemples plus précis seront donnés par après.

La classification permet aussi, selon l'appartenance d'un individu à une classe ou à une autre (par exemple la classe des mammifères) d'en préciser les caractéristiques (par exemple, pour les mammifères, ils allaitent leurs petits), le comportement (par exemple pour une campagne électorale, une campagne publicitaire, ...), etc.

Le problème paraît assez simple mais les difficultés rencontrées sont nombreuses.Tout d'abord, il faut se rendre compte que, contrairement aux ensembles de données tests où l'on cherche bien souvent les classes naturelles (voir chapitre 2), dans les applications réelles, il n'existe pas de classification

à priori. C'est-à-dire que les classes sont inconnues, ainsi que leur nombre et leurs effectifs. Il n'existe donc pas "une bonne classification" et il est alors difficile de juger si une classification obtenue est acceptable. Le deuxième problème vient de la multiplicité des variables en jeu. Il est évident que deux individus, même s'ils se ressemblent beaucoup, n'auront pas nécessairement toutes leurs variables avec des valeurs plus ou moins proches. De plus, il faut aussi savoir repérer quelles sont les variables les plus discriminantes pour pouvoir choisir celles sur lesquelles on va travailler. Enfin, une autre grosse difficulté vient du nombre de méthodes de classification connues. En effet, toutes ces methodes de classification fonctionnent très différemment et ont chacune des propriétés qui leurs sont propres. Le vrai problème est alors qu'aucune de ces méthodes ne "marche" à chaque fois, sur tous les exemples. Donc, il n'y a pas de méthode remède, qui donnerait à chaque fois un meilleur resultat que toutes les autres. Il faut savoir analyser...

Remarquons encore que la classification est une méthode descriptive où l'utilisateur fournit le minimum d'hypothèses sur les données et où toutes les variables jouent le même rôle. Enfin le processus de classification comporte trois grandes étapes:

1. l'obtention des données.
2. la constitution des groupes.
3. l'analyse des résultats.

Le problème auquel nous allons nous intéresser dans ce mémoire fait partie de la constitution des groupes. Il s'agit de la détermination du nombre de groupes.
On ne parlera donc pas de l'obtention des données. Par contre, l'analyse des résultats jouera aussi un rôle important en vue de confirmer ou d'infirmer les résultats obtenus[1].

> "Toutes les connaissances que nous possédons dépendent des méthodes par lesquelles nous distinguons le semblable du dissemblable. Plus le nombre de distinctions que permettent ces méthodes est grand, plus notre connaissance des choses s'accroit. Plus les objets qui nous intéressent sont nombreux, plus il est difficile de créer une telle méthode mais plus cela est nécessaire"
> ([24]).

1. Parfois, la détermination du nombre de groupes est considérée comme appartenant à l'analyse des résultats (de la classification).

1.2 Quelques exemples

1. Le responsable du service de documentation d'une université s'efforce de mettre sur pied une classification systématique des ouvrages et des périodiques disponibles.

2. Le géographe désire effectuer une régionalisation et pour cela veut mettre en évidence les groupes de districts dont le comportement est identique sur un ensemble de facteurs.

3. Le responsable d'une agence de publicité veut segmenter les lecteurs de magazines pour mettre au point une campagne publicitaire adaptée à chaque groupe d'individus.

4. Le candidat politique désireux de fixer sa stratégie électorale définit les divers types d'électeurs habitant sa circonscription.

5. Le psychiatre, le medecin, essaient de mettre en évidence des familles de maladies avec pour objectif d'adapter le traitement à chacun des types de maladies.

1.3 La constitution des groupes

1.3.1 Introduction au problème

Soit un ensemble de n **individus** :

$$E = \{x_1, x_2, ..., x_n\}.$$

Chaque individu est caractérisé par un ensemble de **p variables** :

$$V = \{v_1, v_2, ..., v_p\}.$$

Cette situation peut se représenter par une matrice $n \times p$, que l'on appellera matrice des données :

$$\mathcal{M} \equiv \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \dots & & x_{np} \end{pmatrix}.$$

Pour pouvoir visualiser ces objets, on a l'habitude de représenter chacun de ces objets par un point dans un espace à p dimensions, chacune de ces dimensions représentant une des p variables descriptives. Remarquons que pour cela, il faut bien entendu que les variables utilisées soient des variables quantitatives (ce que nous supposerons dans ce mémoire).
Chaque point i représentant l'objet x_i a alors pour coordonnées :

$$(x_{i1}, x_{i2}, ..., x_{ip}).$$

On se rappelle alors que le but est de trouver des classes de façon à ce que les membres de chaque classe aient en commun certaines caractéristiques qui les distinguent des membres des autres classes (on peut alors essayer de coller une "étiquette" sur chacune des classes). Mais comment faire ? C'est ce que nous verrons dans les paragraphes suivants.

N.B.: Pour p valant 2 ou 3, cela peut se faire de façon visuelle.

1.3.2 Comment mesurer la proximité entre deux individus ?

A chaque paire d'objets x_i et x_j, on va associer un nombre que l'on appellera indice de similarité ou de dissimilarité selon le cas, et qui mesure la proximité entre deux objets (pour pouvoir voir à quel point ils sont semblables ou différents).

Dans ce mémoire, la proximité entre deux objets sera mesurée par la distance euclidienne non-pondérée :

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \ .$$

C'est un indice de dissimilarité car, vu la représentation adoptée au paragraphe précédent, plus la distance entre deux objets est grande, plus ils sont "dissemblables".
On peut ainsi définir la matrice des proximités :

$$\mathcal{D} \equiv \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & & \ddots & \vdots \\ d_{n1} & \dots & & d_{nn} \end{pmatrix} .$$

Remarques :

- Cette distance dépend de l'unité de mesure et de la variance de chaque variable.

- Lorsque les unités de mesure diffèrent d'une variable à l'autre, la variable ayant la plus forte variance prendra une importance prépondérante dans la distance entre les objets. La distance euclidienne est donc biaisée en direction des variables qui ont la plus grande dispersion. Afin de pallier à ces difficultés, on procède couramment à la standardisation des variables en soustrayant à chaque mesure la moyenne de la variable et en divisant par l'écart type. Toutes les variables seront ainsi mesurées en unité d'écart-type. En conséquence, les distances entre objets sont elles-mêmes mesurées en écart-type. De plus, la standardisation des variables ne préserve pas l'ordre des distances obtenu sur les données brutes. Remarquons enfin que la standardisation peut entrainer une perte d'information pas nécessairement bénéfique à l'analyse. Il faut alors bien réfléchir pour voir si la standardisation est (ou était si on analyse les résultats) pertinente.

- On peut aussi choisir d'accorder un poids différent à chaque variable, indépendamment de la standardisation (par exemple, selon la qualité ou la fiabilité des mesures effectuées) :

$$d_{ij}^2 = \sum_{k=1}^{p} p_k (x_{ik} - x_{jk})^2 \ .$$

1.3.3 Position du problème

Reprenons notre ensemble d'individus :

$$E = \{x_1, x_2, ..., x_n\}.$$

On recherche une partition[1] de E en k classes (k fixé) :

$$P = \{C_1, C_2, ..., C_k\}.$$

A chaque partition P, on peut associer un critère de classification :

$$\begin{array}{rccc} W: & P_k & \longrightarrow & R^+ \\ & P & \rightsquigarrow & W(P,k). \end{array}$$

En général, le problème est alors :

Trouver la partition "optimale",

$$P^* = \{C_1^*, C_2^*, ..., C_k^*\} \quad \textbf{telle que} \quad W(P^*,k) = \min_{P \in P_k} W(P,k)$$

ou

$$W(P^*,k) = \max_{P \in P_k} W(P,k)$$

Ce problème peut avoir différentes facettes :

- Soit k est connu et il faut alors trouver la partition qui donne une valeur de $W(P,k)$ optimale.
- Soit k est inconnu et est également à déterminer pour que le critère soit optimal[2]. **C'est le cas auquel nous nous intéresserons.**

Exemple de critère :
La somme des carrés des distances entre les points à l'intérieur des groupes. On cherche alors la valeur minimale de ce critère. De plus, ce critère reflète un désir de trouver des ensembles sphériques de variance minimale.

1. Voir annexes.
2. Beaucoup de critères dépendent de k.

1.3.4 Difficulté liée à ce problème et solutions

Si on appelle $S(n,k)$ le nombre total de partitions $P = \{C_1, C_2, ..., C_k\}$ de n objets en k classes, alors $S(n,k)$ vaut :

$$S(n,k) = \frac{1}{k}\sum_{i=1}^{k} C_n^i (-1)^{k-i} i^n$$

avec $S(n,k) = kS(n-1,k) + S(n-1,k-1)$
$S(1,1) = 1$; $S(1,n) = 0$ $n \neq 1$
(nombre de stirling d'ordre 1)
([1])

Voyons ce que cela donne sur quelques exemples :

$S(n,2) = 2^{n-1} - 1 \Rightarrow S(59,2) \approx 10^{18}$
$S(15,3) = 2.375.101$
$S(20,4) = 45.232.115.901$
$S(100,5) \approx 10^{68}$

En admettant qu'un ordinateur évalue un million de partitions par seconde, il faudrait alors huit jours pour calculer les $S(20,5)$ partitions possibles de 20 objets en 5 groupes et déjà deux mille quatre cents quarante quatre siecles pour calculer les $S(30,5)$ partitions possibles de 30 objets en 5 groupes. Comme de plus le nombre de classes est à priori inconnu, le nombre total de partitions, encore appelé nombre de Bell, est:

$$B(n) = \sum_{k=1}^{n} S(n,k) = e^{-1} \sum_{k=1}^{\infty} \frac{k^n}{k!}$$

Par exemple, pour n=15, $B(n)$ vaut 1.382.958.545 .

Il y a alors deux façons de contourner ce problème:

1. Les diverses méthodes de classification qui, par des processus particuliers, essaient de ne s'intéresser qu'aux "meilleures" partitions possibles (voir point 5). Ces méthodes trouvent alors des solutions locales.

2. L'algorithme de Fisher[1] (voir point 7).

1. Il se limite au cas où il n'y a qu'une seule variable.

1.3.5 Classement des méthodes de classification

N.B.: Ne seront détaillées que les classes de méthodes utilisées par la suite.

Tout d'abord, il faut savoir qu'il existe deux grandes classes de méthodes :

1. Les méthodes monothétiques.
   Leur objectif est la recherche d'une hiérarchie de partitions [1] , construite à partir de la matrice des données, par une suite de divisions en deux classes ne tenant compte que d'une seule variable à la fois. A chaque étape, la division peut se faire suivant une variable différente.
   Dans ces méthodes, une classe d'objets est définie par la possession en commun d'un attribut. Par exemple, on peut désirer classer ensembles tous les individus qui ont répondu de la même façon à l'une des questions d'une enquête. Le problème se pose alors de sélectionner la question la plus discriminante ou la plus sélective, c'est-à-dire celle qui apporte le plus d'informations sur l'ensemble des réponses.

2. Les méthodes polythétiques.
   Les méthodes polythétiques sont celles auxquelles on va s'intéresser dans ce mémoire. Elles tiennent compte simultanément de l'ensemble des variables décrivant les objets. Ainsi, deux objets pourront appartenir à la même classe sans posséder un seul caractère commun pourvu qu'ils se ressemblent suffisamment du point de vue de l'indice de similarité choisi pour mesurer leur ressemblance. L'information de base manipulée par les méthodes polythétiques est la matrice des proximités et non pas la matrice des données.

   Ces méthodes peuvent être conçues, ainsi que le remarquent Jardine et Sibson ([23]), comme transformant la matrice des proximités en une nouvelle matrice dans laquelle les groupes d'objets sont plus apparents que dans la matrice des proximités initiale. La transformation remplace les dissimilarités initiales par de nouvelles dissimilarités vérifiant des propriétés plus fortes.

0. Voir annexes.

Les méthodes hiérarchiques

Leur objectif est la recherche d'une famille de partitions[1] telle que les groupements ou les divisions successifs des objets forment une hiérarchie.

Deux cas sont alors possibles :

1. Algorithmes agglomératifs.
   - on part de n classes : $C_1 = \{x_1\}, C_2 = \{x_2\}, ..., C_n = \{x_n\}$.
   - à chaque étape, on regroupe les deux classes les "plus proches" (pour avoir dans une même classe des objets qui se ressemblent)

     $\rightarrow$ étape 0 : n classes
     étape 1 : n-1 classes
     $\vdots$
     étape n-1 : 1 classe $\equiv E = \{x_1, x_2, ..., x_n\}$.
   - pour chaque définition différente de la distance entre deux classes, on aura une méthode différente (à chaque étape, les deux classes fusionnées seront différentes).
2. Algorithmes divisifs.
   - on part d'une classe : $E = \{x_1, x_2, ..., x_n\}$.
   - à chaque étape, on va choisir parmi toutes les divisions possibles d'une des classes en deux, celle dont la distance entre les classes obtenues par cette division est maximale

     $\rightarrow$ étape 0 : 1 classe
     étape 1 : 2 classes
     $\vdots$
     étape n-1 : n classes .
   - à nouveau, pour chaque définition différente de la distance entre deux classes, on aura une méthode différente.

1. Voi annexes.

Enfin, il existe différentes classes de méthodes hiérarchiques:

1. Les méthodes hiérarchiques ordinales.
   Elles n'utilisent pas d'autre information que le classement de paires d'objets par ordre de proximité.

2. Les méthodes hiérarchiques non-ordinales.
   Ces méthodes, contrairement aux précédentes, utilisent les valeurs numériques des dissimilarités entre paires d'objets.

Remarques :

- La hiérarchie peut dépendre de l'ordre d'introduction des données.

- Le problème avec les méthodes hiéarchiques est qu'une fois qu'un individu est mis dans un groupe, il y restera jusque la fin. En d'autres mots, il n'est pas possible de corriger une mauvaise partition (par exemple, pour une méthode ascendante, si deux objets sont dans le même groupe à une étape i, pour toute étape $j \geq i$, ils resteront dans ce même groupe).

- Il existe aussi des méthodes hiérarchiques générant une hiérarchie de recouvrements (ordonnés par la relation de finesse) au lieu d'une hiérarchie de partitions.

Les méthodes de réallocation

Ces méthodes ont pour but de construire une partition unique des objets en k classes où le nombre k est soit spécifié à priori, soit determiné par l'algorithme. L'idée centrale de ces méthodes est de choisir une partition initiale des objets et de déplacer les objets d'un groupe à l'autre pour obtenir une partition meilleure. Les nombreux algorithmes existants diffèrent par le choix de la partition initiale, par la définition qu'ils donnent à une "meilleure partition" et par la méthode utilisée pour améliorer la partition. Ils partent donc d'une partition initiale, puis génèrent un ensemble de partitions successives permettant d'améliorer la valeur d'une fonction objective (d'un critère) jusqu'à ce qu'un minimum soit atteint. Ces méthodes sont en général simples et économiques.

Les méthodes de recherche de densité

Comme on considère les objets comme des points dans un espace à p dimensions, il est naturel de penser aux groupes d'objets en terme de régions de l'espace où la densité de points est grande, séparées par des régions où elle ne l'est pas. De façon générale, on recherche des régions de forte densité pour constituer des groupes.

1.3.6 Partition optimale par l'algorithme de Fischer

Comme on l'a déjà vu, la découverte d'une partition optimale par simple énumeration est impossible dès que n dépasse une quinzaine d'objets (15 objets : 1,4 milliard de partitions). Cependant, lorsque les objets sont ordonnés, par exemple dans le temps, et que l'on impose aux classes d'être formées d'objets contigus, le nombre de partitions en k classes contigües n'est plus que de C_{n-1}^{k-1}. Il devient alors possible de découvrir la partition optimisant un critère d'homogénéité des classes (alors que beaucoup d'algorithmes conduisent à une solution approchée).

1.3.7 Présentation des méthodes de classification utilisées

Elles sont au nombre de cinq. Ce sont toutes des méthodes hiérarchiques sauf la méthode des hypervolumes dont nous vous parlerons en dernier. Pour les quatre méthodes hiérarchiques, il existe à la fois des algorithmes ascendants et descendants, mais nous n'utiliserons que les algorithmes ascendants. Rappelons que pour les méthodes hiérarchiques ascendantes, nous partons d'une partition où chaque élément forme une classe, et à chaque étape, nous fusionnons les deux classes les plus proches.
La seule chose qui distinguera donc ces différentes méthodes hiérarchiques sera la façon de mesurer la distance entre deux groupes d'objets.

La méthode du voisin le plus proche

1. Description :

   Cette méthode mesure la distance entre deux classes C_i et C_j d'une partition par la plus petite distance séparant un point d'une classe et un point de l'autre :

   $$d_{C_iC_j} = \min_{x \in C_i, y \in C_j} d(x, y).$$

   On appelle cette distance la distance du saut minimum.

   Exemple:

1.	2.	3.			
4.	5.	6.			
7.	8.	9.	.10	.11	.12
			.13	.14	.15
			.16	.17	.18

   Si $C_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

   et $C_2 = \{10, 11, 12, 13, 14, 15, 16, 17, 18\}$

   alors $d_{C_1C_2} = d(9, 10)$.

2. Exemple :

$$E = \{1, 2, 3, 4\}$$

Voici la matrice des distances entre ces quatre objets :

$$D_0 = \begin{pmatrix} 0 & 5 & 9 & 8 \\ 5 & 0 & 4 & 5 \\ 9 & 4 & 0 & 3 \\ 8 & 5 & 3 & 0 \end{pmatrix}$$

-Etape 0 : $P_0 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

-Etape 1 : Calculons les distances entre chaque paire de classes.
Comme toutes les classes sont constituées d'un point, la distance les séparant est la distance entre ces points !
On a $d_{12} = 5$, $d_{13} = 9$, $d_{14} = 8$, $d_{23} = 4$, $d_{24} = 5$, $d_{34} = 3$.
Les deux points les plus proches sont 3 et 4.
Nous allons donc les regrouper:

$P_1 = \{\{1\}, \{2\}, \{3, 4\}\}$

-Etape 2 : Calculons les distances entre chaque paire de classes.
$d_{12} = 5$
$d_{1\{3,4\}} = \min\{d_{13}, d_{14}\} = \min\{9, 8\} = 8$
$d_{2\{3,4\}} = \min\{d_{23}, d_{24}\} = \min\{4, 5\} = 4$
On regroupe les deux classes les plus proches :
$\{2\}$ et $\{3, 4\}$.

$P_2 = \{\{1\}, \{2, 3, 4\}\}$

-Etape 3 : Le seul groupement possible est :
$\{1\}$ et $\{2, 3, 4\}$.

$P_3 = \{\{1, 2, 3, 4\}\}$

3. Propriétés :

   (a) La hiérarchie peut dépendre de l'ordre de lecture des données (par exemple, si la distance entre deux objets est la même que la distance entre deux autres objets, l'algorithme choisira comme distance minimale la première qu'il a calculée).

   (b) La méthode est peu robuste car en perturbant un peu les données, on peut modifier beaucoup la hiérarchie obtenue.
   Cela est dû au fait que l'on mesure la distance entre deux groupes par la seule plus petite distance les séparant.

   (c) A chaque fusion, les objets non encore classés tendent à être incorporés aux groupes existants plutôt qu'à former de nouveaux groupes ([19]). En conséquence, la méthode donne des résultats peu satisfaisants lorsque des objets intermédiaires sont présents entre deux groupes ou lorsque les groupes ne sont pas nettement séparés (propriété de chaînage).

La méthode du voisin le plus éloigné

1. Description :

   Cette méthode mesure la distance entre deux classes C_i et C_j d'une partition par la plus grande distance séparant un point d'une classe avec un point de l'autre :

   $$d_{C_i,C_j} = \max_{x \in C_i, y \in C_j} d(x,y).$$

   On appelle cette distance la distance du saut maximum.

   Exemple:

1.	2.	3.			
4.	5.	6.	.10	.11	.12
7.	8.	9.	.13	.14	.15
			.16	.17	.18

   Si $C_1 = \{1,2,3,4,5,6,7,8,9\}$

   et $C_2 = \{10,11,12,13,14,15,16,17,18\}$

   alors $d_{C_1C_2} = d(1,18)$.

   ATTENTION : pour calculer la distance entre deux classes, on prend la distance **maximale** entre deux éléments (un de chaque classe), mais ensuite, on regroupe à nouveau les deux classes dont la distance est **minimale.**

2. Exemple :

$$E = \{1, 2, 3, 4\}$$

Voici la matrice des distances entre ces quatre objets :

$$D_0 = \begin{pmatrix} 0 & 6 & 10 & 9 \\ 6 & 0 & 4 & 5 \\ 10 & 4 & 0 & 3 \\ 9 & 5 & 3 & 0 \end{pmatrix}$$

-Etape 0 : $P_0 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

-Etape 1 : Calculons les distances entre chaque paire de classes.
Comme toutes les classes sont constituées d'un point, la distance les séparant est la distance entre ces points !
On a $d_{12} = 6$, $d_{13} = 10$, $d_{14} = 9$, $d_{23} = 4$, $d_{24} = 5$, $d_{34} = 3$.
Les deux points les plus proches sont 3 et 4.
Nous allons donc les regrouper:

$P_1 = \{\{1\}, \{2\}, \{3, 4\}\}$

-Etape 2 : Calculons les distances entre chaque paire de classes.
$d_{12} = 6$
$d_{1\{3,4\}} = \max\{d_{13}, d_{14}\} = \max\{6, 10\} = 10$
$d_{2\{3,4\}} = \max\{d_{23}, d_{24}\} = \max\{4, 5\} = 5$
On regroupe les deux classes les plus proches :
$\{2\}$ et $\{3, 4\}$.

$P_2 = \{\{1\}, \{2, 3, 4\}\}$

-Etape 3 : Le seul groupement possible est :
$\{1\}$ et $\{2, 3, 4\}$.

$P_3 = \{\{1, 2, 3, 4\}\}$

3. Propriétés :

   (a) L'algorithme ascendant et l'algorithme descendant ne fournissent pas toujours la même hiérarchie.

   (b) La hiérarchie dépend de l'ordre de lecture des données.

   (c) La méthode est peu robuste.

   (d) La méthode a tendance à former des groupes hypersphériques.

La méthode de la moyenne (group average)

1. Description :

   Cette méthode mesure la distance entre deux classes C_i et C_j comportant respectivement n_i et n_j objets, par la valeur moyenne des distances inter-classes :

   $$d_{C_i,C_j} = \sum_{x \in C_i, y \in C_j} d(x,y)/n_i n_j$$

2. Propriétés :

   (a) Cette méthode ne garantit pas que les classes soient le plus homogène possible. Elle ne garantit pas que les deux classes fusionnées soient composées d'objets proches en moyenne. Cela est du au fait que le critère de la moyenne ne fait pas intervenir les dissimilarités intra-classes.

   (b) Remarque : les groupes obtenus diffèrent peu, en général, de ceux obtenus par le critère du voisin le plus éloigné.

La méthode de Ward

1. Description :

   La distance entre deux classes C_i et C_j comportant respectivement n_i et n_j objets, est mesurée par la différence entre le moment centré d'ordre 2 des classes fusionnées et le moment centré d'ordre 2 de chacune des classes :

   $$d^2_{C_iC_j} = M^2(C_i \cup C_j) - M^2(C_i) - M^2(C_j)$$

   où $M^2(C_i) = \sum_{i \in C_i} \sum_{k=1}^{p} (x_{ik} - \overline{x}_k)^2 = \sum_{i,j \in C_i} d_{ij}/n_i$

   $\equiv$ somme des carrés des écarts des objets au centre de gravité de la classe.

   $\equiv$ moyenne des distances euclidiennes entre toutes les paires d'objets de la classe.

   Il est facile de démontrer ([22]) que ce critère du moment centré d'ordre 2 se ramène à une pondération de la distance entre centres de gravité :

   $$d_{C_iC_j} = \left(\frac{n_i n_j}{n_i + n_j} \right)^{\frac{1}{2}} \|g(C_i) - g(C_j)\|$$

   où $g(C_i)$ et $g(C_j)$ sont les centres de gravité des classes C_i et C_j respectivement.

   En fait, on fusionnne les deux groupes qui conduisent à un accroissement minimum du critère des moindres carrés.

2. Propriétés :

   (a) Croissance monotone des distances à chaque fusion.

   (b) Cette méthode tend aussi à former des groupes hypersphériques (cela est dû au fait qu'elle est basée sur le critère des moindres carrés).

La méthode des hypervolumes

- Description :

Reprenons la formulation de départ du problème de classification. Nous avons une valeur du critère de classification $W(P,k)$ pour chaque partition P en k classes:

$$\begin{array}{rccc} W: & P_k & \longrightarrow & R^+ \\ & P & \rightsquigarrow & W(P,k). \end{array}$$

Le problème est alors d'en trouver l'optimum.
Dans la suite de ce paragraphe, nous allons expliquer le critère des hypervolumes.

1. Problème initial : estimation d'un domaine convexe.
   Soit le problème suivant, proposé par D.G. Kendall: "Etant donné la réalisation d'un processus de Poisson homogène d'intensité inconnue à l'intérieur d'un ensemble convexe compact D, trouver D en utilisant des méthodes d'inférence statistique". La solution de ce problème fut trouvée par Ripley et Rasson ([30]) et Rasson ([28]).

   L'estimation proposée est $D' = g(H(x)) + cs(H(x))$ où
   - $x = \{x_1, x_2, ..., x_n\}$ est la réalisation du processus de Poisson dans D;
   - H(x) est l'enveloppe convexe de x;
   - g(H(x)) est le centroide de H(x);
   - $s(H(x)) = H(x) - g(H(x))$.

   Il s'agit donc d'une expansion homothétique de l'enveloppe convexe à partir de son centre de gravité. La constante de dilatation c peut être estimée par $c = \sqrt{\frac{n}{n-v_n}}$ où v_n est le nombre de sommets de H(x) ([26]).

2. Application en classification.
   Maintenant, supposons que les points que nous observons sont générés par un processus de Poisson homogène dans $D \subset R^n$ où:
   - $D = \cup_{i=1}^{k} D_i$.
   - les ensembles D_i sont disjoints.

Notons:

- $x = (x_1, x_2, \ldots, x_n)$ la réalisation du processus dans D.
- $C_i \subset \{x_1, x_2, \ldots, x_n\}$, l'ensemble des observations appartenant à D_i $(1 \leq i \leq k)$.

Le problème est alors d'estimer les domaines inconnus D_i dans lesquels les points sont distribués. Si nous utilisons la méthode du maximum de vraisemblance pour estimer ce domaine, nous trouvons alors que: *le domaine D pour lequel la vraisemblance sera maximale est, parmi ceux qui contiennent tous les points, celui dont la mesure de Lebesgue est minimale* ([15];[16]).
Pour que ce problème soit bien posé statistiquement, on doit imposer la convexité des ensembles D_i et le critère correspondant, appelé critère des hypervolumes, s'écrira comme suit :

$$\begin{array}{rccl} W: & P_k & \longrightarrow & R^+ \\ & P & \rightsquigarrow & W(P,K) = \sum_{i=1}^{k} m(H(C_i)) \end{array}$$

où :

- $H(C_i)$ est l'enveloppe convexe des points appartenant à C_i.
- $m(H(C_i))$ est la mesure de Lebesgue multidimensionelle de cette enveloppe convexe.

Dans R, par exemple, on recherche la partition qui minimise la somme des longueurs des enveloppes convexes des points se trouvant dans chacune des classes. Dans R^2 ou R^3, on recherche respectivement les partitions qui minimisent la somme des aires ou des volumes de ces enveloppes convexes. Enfin, dans R^m, on cherche à rendre minimale la somme des hypervolumes de ces enveloppes convexes.

En ce qui concerne la recherche de la partition optimale, il existe deux algorithmes :

- Le premier est un algorithme global. Il utilise la technique du "branch and bound" et permet de trouver les solutions pour des valeurs de n et k petites (le temps de calcul étant assez élevé).
- Le deuxième est un algorithme local. Son temps calcul est assez élevé aussi. Notons que si il y a un "espace vide" entre les différentes classes, il trouve aussi la solution optimale.

1.3.8 Calcul pratique des distances pour les méthodes utilisées

En fait, les distances entre groupes pour les quatre méthodes utilisées satisfont une relation de récurrence ([11]).
Quand on veut calculer la distance entre un groupe C_k et un groupe C_{ij} formé par la fusion des groupes C_i et C_j (ce qui est le cas à chaque étape d'une méthode hiérarchique ascendante), la formule suivante est vérifiée:

$$d_{C_kC_{ij}} = \alpha_i d_{C_kC_i} + \alpha_j d_{C_kC_j} + \beta d_{C_iC_j} + \gamma |d_{C_kC_i} - d_{C_kC_j}|$$

où α_i, α_j, β et γ sont des paramètres dont les valeurs changent selon la méthode utilisée.

1. Voisin le plus proche: $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = 0$; $\gamma = -\frac{1}{2}$.

2. Voisin le plus éloigné: $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = 0$; $\gamma = \frac{1}{2}$.

3. Méthode de la moyenne: $\alpha_i = \frac{n_i}{n_i+n_j}$; $\alpha_j = \frac{n_j}{n_i+n_j}$; $\beta = \gamma = 0$.

4. Ward: $\alpha_i = \frac{n_k+n_i}{n_k+n_i+n_j}$; $\alpha_j = \frac{n_k+n_j}{n_k+n_i+n_j}$; $\beta = \frac{-n_k}{n_k+n_i+n_j}$; $\gamma = 0$.

La formule de récurrence ci-dessus a été utilisée pour la programmation des méthodes de classification utilisées car elle est très facile à implémenter.

Chapitre 2

La validation des résultats : détermination du nombre de classes

2.1 Motivation : des problèmes divers...

2.1.1 Introduction

Comme cela a été signalé au chapitre premier, le problème auquel nous allons nous intéresser est la détermination du nombre de classes. Mais, il faut savoir que les classifications[1] auxquelles nous allons appliquer ces méthodes ne sont pas toujours celles attendues[2] pour toutes une série de raisons. Pour parer à ce problème et donc pour pouvoir valider les résultats, il faut prendre en compte le plus de sources d'erreurs possible. Car en effet, les difficultés peuvent provenir directement des méthodes de classification comme elles peuvent provenir de leur utilisation combinée avec les méthodes de détermination du nombre de classes, etc. C'est aussi pourquoi, avant d'expliquer comment nous avons travaillé et dans quel but, nous allons nous attacher à définir les divers problèmes existants en classification.

1. Ces classifications étant obtenues par une méthode de classification quelconque.

2. Celles où les classes trouvées sont les classes "naturelles" dans notre cas (la notion de classe naturelle sera définie à la page suivante).

2.1.2 Les problèmes rencontrés

La notion de classe naturelle

Le problème de base de toute classification étant de trouver des classes parmi un ensemble d'objets, la première chose à faire est de bien définir ce que nous allons considérer comme une classe.
Commençons par nous rappeler ce que l'on sait déjà:
les classes sont des ensembles d'individus qui ont des caractéristiques en commun qui les distinguent des individus des autres classes.
Nous avons aussi vu que nous représentions les individus par des points dans l'espace, les individus se ressemblant étant à une petite distance euclidienne les uns des autres. Nous pouvons maintenant affirmer que:
les classes sont des groupes de points dans l'espace de travail tels que, "en général", la distance entre deux points d'une même classe est plus petite que la distance entre deux points appartenant à deux classes différentes.
Nous allons voir pourquoi ce "en général".

Prenons par exemple ces groupes de points:

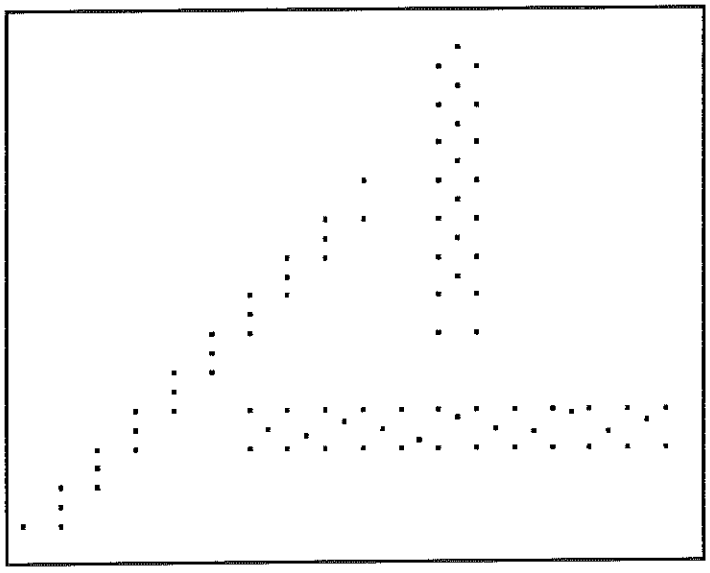


Figure 2.1.

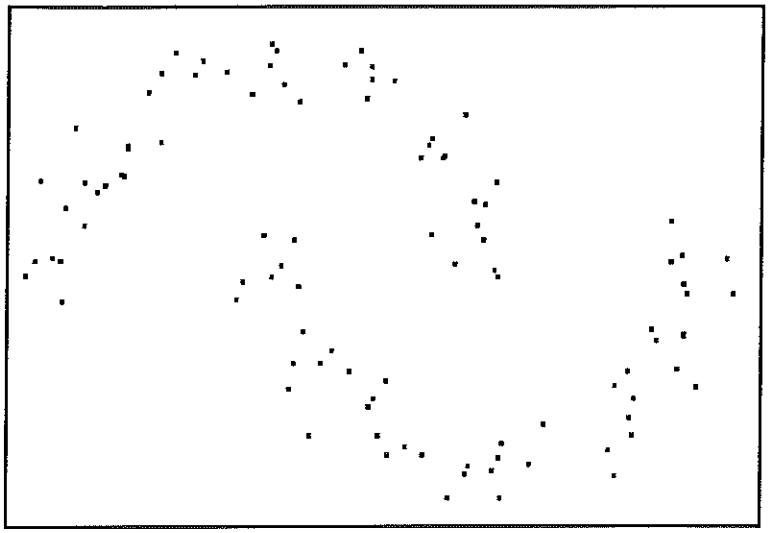


Figure 2.2.

Quels sont les classes présentes?
Correspondent-elles à notre définition?

Nous allons définir les choses d'une façon un peu plus intuitive mais de telle manière qu'aucun cas ne pose problème.
Comme cela a été signalé, dans ce travail, nous allons rechercher des classes "naturelles". **Lorsqu'on mesure sur chaque individu deux variables quantitatives, il s'agit des classes que l'oeil repère lorsqu'on représente les individus comme des points dans un espace à deux dimensions. Ces notions se généralisent facilement dans des espaces de dimension supérieure.**
Nous appellerons donc ces classes les **classes naturelles** et nous jugerons qu'**une classification donne de bons résultats si elle retrouve les classes naturelles.**

Prenons un exemple:

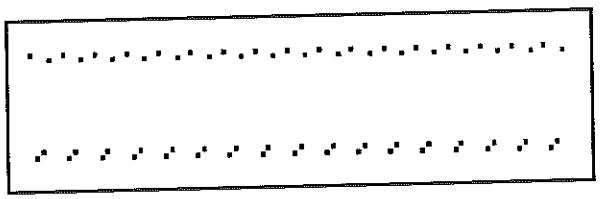


Figure 2.3.

Si nous appliquons la méthode du voisin le plus éloigné à cet ensemble de points, nous retrouverons comme classes :

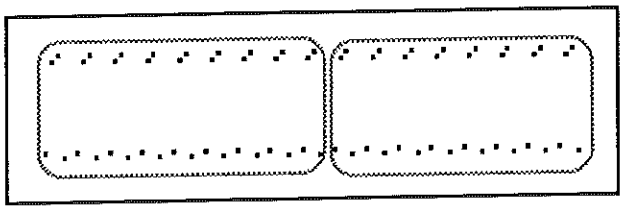


Figure 2.4.

Tandis que si nous appliquons la méthode du voisin le plus proche, nous aurons :

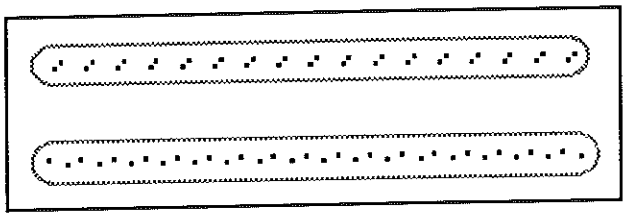


Figure 2.5.

Dans notre cas, nous jugerons que la méthode du voisin le plus éloigné n'a pas donné de bons résultats tandis que la méthode du voisin le plus proche a donné exactement les résultats attendus.

Les exemples que nous considérerons sont pour la plupart des ensembles de données test que l'on trouve dans la littérature scientifique et qui servent bien souvent à tester des nouvelles méthodes de classification ou de détermination du nombre de classe. **Ces exemples présentent souvent des classes naturelles évidentes.**

Stabilité des méthodes de classification

Comme nous l'avons vu au premier chapitre, toutes les méthodes de classification ont leurs propres biais. Par exemple, la méthode du voisin le plus éloigné privilégie les classes hyperspériques. Elle ne retrouve donc pas toujours la structure naturelle des données. Signalons que ces biais[1] ont été en général trouvés en appliquant les diverses méthodes à des "données test" avec des caractéristiques bien spécifiques (comme celles du paragraphe précédent).

Enfin, outre ces biais, des conditions d'admissibilité ont été développées pour tester ces méthodes et permettent de les caractériser. Ont ainsi été définies : l'admissibilité par convexité, par connectivité, par rapport aux images, par rapport aux proportions des classes, par omission de classes, et par répétabilité ([12]; [32]). On dit par exemple qu'une procédure est admissible par rapport aux proportions des classes si une duplication de chaque classe un nombre arbitraire de fois ne modifie pas les frontières des classes.
Ces propriétés sont assez importantes et nous verrons dans le chapitre 4 que certaines méthodes de détermination du nombre de classes n'y sont pas insensibles non plus. Par exemple, certains indices varient et ne donnent plus de bons résultats si une classe comporte plus de points qu'une autre. Ce qui est bien sur lié à l'admissibilité par rapport aux proportions des classes.

Enfin, signalons qu'une dernière propriété peut être intéressante. Il s'agit de la robustesse d'une méthode. Elle se définit comme la capacité à générer des classifications stables[2].

1. Ces biais ont été signalés lors de la présentation des méthodes au premier chapitre.

2. Et cela dans le sens suivant : si on perturbe un peu les données et qu'on applique à nouveau la méthode de classification, cela ne change pas la classification obtenue.

Problèmes liés aux méthodes de détermination du nombre de classes

Une première chose à remarquer est que malgré qu'il y ait beaucoup de littérature sur les différents indices existant pour trouver le nombre de classes, très peu d'articles ou de livres ont comparé ces différents indices. De plus, les biais de ceux-ci sont rarement connus. Il est donc important de se méfier des résultats obtenus et de bien les analyser. Il est d'ailleurs ainsi parfois possible de détecter les caractéristiques de ces méthodes [1].

Conclusion

Il n'est pas facile de retrouver les classes désirées, c'est-à-dire dans notre cas les classes naturelles !

1. Voir chapitre 4.

2.2 Comment avons-nous travaillé

2.2.1 Travaux de base

Tout d'abord, il faut savoir que plusieurs travaux (articles) avaient déjà été réalisés afin de comparer les méthodes existant pour déterminer le nombre de classes et essayer de trouver quelles étaient les meilleures de ces méthodes. Parmi ceux-ci, nous nous sommes servi de deux pour la réalisation de ce mémoire. Nous allons examiner un peu la façon dont ils procédaient.

Le travail de A. Hardy ([18]):

Le but de ce travail était de comparer trois méthodes de détermination du nombre de classes basées sur le critère des hypervolumes (développé à Namur par A. Hardy et J.P. Rasson) avec quatre autres méthodes plus connues. Le critère des hypervolumes ainsi que les méthodes de détermination du nombre de classes seront expliqués dans le troisième chapitre. En attendant, voici un petit résumé de la démarche suivie:

- Tout d'abord, six ensembles de points en deux dimensions et présentant des particularités intéressantes[1] ont été choisis.
- Ensuite, six méthodes de classification différentes[2] ont été appliquées à chacun de ces ensembles de points.
- Enfin, pour chaque ensemble de points et pour chacun des résultats obtenus par les méthodes de classification, les sept méthodes de détermination du nombre de classes ont été appliquées.

On pouvait ainsi voir quelles méthodes retrouvaient le plus souvent le bon nombre de classes.

De plus, les exemples choisis ayant des propriétés particulières quand on les représentes dans le plan (classes allongées, ...), cela permettait de dégager quelques axes de réflexion quant aux propriétés des méthodes de détermination du nombre de classes. Par exemple, si une méthode ne marche dans aucun exemple où les classes sont allongées, on peut voir cela comme une "caractéristique" de cette méthode.

Ensuite, notons que les quatre méthodes de détermination du nombre de classes non basées sur le critére des hypervolumes ont été choisies en raison

1. Où les classes naturelles étaient évidentes.

2. Car en employant des méthodes différentes, les résultats ne dépendent pas d'une méthode. De plus, on voit avec quelles méthodes un indice donne les meilleurs résultats.

de leur disponibilité dans le logiciel Clustan qui est le logiciel le plus complet et le plus développé en classification automatique.

Enfin, signalons que les méthodes basées sur les hypervolumes donnaient dans presque tous les cas de meilleurs résultats que les autres.

Le travail de Milligan et Cooper ([25]) :

Dans ce travail, Milligan et Cooper ont procédé de manière différente. Tout d'abord, leur but était de comparer trente méthodes de détermination du nombre de classes. Ensuite, au lieu de choisir quelques ensembles de points bien particuliers, Milligan et Cooper ont pris beaucoup d'ensembles de points mais générés de façon aléatoire. Chaque ensemble de points contenait 2, 3, 4 ou 5 groupes différents, ne se recouvrant pas[1], avec un total de cinquante points chacun et dans un espace à 2, 4 ou 6 dimensions. A nouveau, pour disposer d'une variété de solutions, les ensembles de points ont été analysés par quatre méthodes de classification (toutes hiérarchiques).

Ce travail a donc permis de classer les trente règles d'arrêt de la meilleure à la moins bonne. La meilleure étant celle qui a donné le plus souvent les résultats attendus. Evidemment, même si ce classement n'est valable que pour les exemples utilisés, il est fort peu probable qu'une méthode classée première après plusieurs centaines d'exemples (108 jeux de données * 4 méthodes) ne s'avère être "mauvaise". Donc, ce classement donne une indication sur les meilleures méthodes parmi les trente méthodes analysées et permet de les comparer. Il est également important de remarquer que le classement reflète les performances des méthodes où la structure des classes est assez "nette". Ce classement peut donc être altéré pour des structures de données rencontrées dans la vie courante. De plus, le fait que les structures des différents groupes étaient assez marquées permet aussi d'affirmer que les méthodes qui donnaient des mauvais résultats dans ce travail n'avaient aucune chances de donner de bons résultats dans des exemples plus complexes.

1. En fait, les groupes étaient bien séparés les uns des autres tout en présentant chacun une certaine cohésion.

Maintenant, décrivons ce travail d'un point de vue plus pratique. Si nous définissons deux types d'erreur:

- le premier : dire qu'il y a k groupes présents dans un ensemble de points alors qu'en réalité, il y en a moins que k.
- le deuxième : dire qu'il y a k groupes dans un ensemble de points alors qu'en réalité, il n'y en a plus que k.

Le deuxième type d'erreur a été considéré comme plus "grave" car de l'information est perdue en fusionnant[1] deux groupes distincts.
De plus, les ensembles de points ont été générés par un processus très particulier (Milligan et Cooper, 1985; Milligan, 1980 ;Milligan, 1981) favorisant la création de classes "naturelles". Sur 432 solutions (108 ensembles de points * 4 méthodes), 400 avaient une structure "nette".
Remarquons encore que les méthodes de détermination du nombre de classes choisies étaient toutes indépendantes des méthodes de classification utilisées et qu'il est possible de les adapter pour être utilisées avec des méthodes de classification non-hiérarchiques.
Signalons aussi que toutes les méthodes où intervenaient la subjectivité humaine pour décider du nombre de classes (comme par exemple certaines méthodes graphiques) ont été évitées. Les seules méthodes qui ont été choisies étaient donc celles où les règles de décision sur le nombre de groupes présents étaient "automatiques".

Enfin, chaque méthode était adaptée pour donner les meilleurs résultats possibles. Par exemple, quand une méthode permettait de choisir soit la valeur maximale de l'indice, soit la différence maximale entre deux valeurs successives de l'indice, celle qui indiquait le bon nombre de groupes était choisie pour représenter le résultat. De plus, pour les tests statistiques, les valeurs des niveaux α étaient parfois assez modérées pour donner des résultats optimaux. Notons également que seules les vingt-cinq dernières valeurs de l'indice (c'est-à-dire celles correspondant à vingt-six ou moins de vingt-six groupes) étaient prises en compte pour déterminer le nombre de classes de l'ensemble de points examiné. Cela est dû au fait que certains indices ont des problèmes quand il y a presque autant de groupes que de points dans l'ensemble de données.

1. On procède par fusions car on est dans le cas de méthodes hiérarchiques.

Les résultats étaient quant à eux présentés sous la forme de tableaux à deux entrées. En voici un exemple pour la méthode de Calinski et Harabasz:

	Number of true clusters				
1.Calinski et Harabasz	2	3	4	5	Overall
2 or fewer	-	-	1	0	1
1 too few	-	12	6	0	18
correct level	96	95	97	102	390
2 too many	3	0	3	6	12
3 or more	5	1	0	0	6

L'entrée dans la deuxième ligne de la troisième colonne indique par exemple que parmi les ensembles de points constitués de quatre groupes, la méthode de Calinski et Harabasz en a retrouvé un en moins, c'est-à-dire trois, à six reprises.

Outre le classement proposé, il est important de signaler que **huit des dix meilleures méthodes présentaient leur plus mauvais résultat quand le nombre de groupes présents était deux**. Ce cas semble être le plus difficile à détecter.

2.2.2 La démarche suivie

L'idée de départ était d'analyser et de comparer les meilleures méthodes de détermination du nombre de classes issues du classement de Milligan et Cooper. A ce sujet, un travail avait déjà été réalisé par Gordon ([14]). Mais comme vous allez le constater, son but et sa façon de procéder diffèrent de la nôtre.
Commençons par examiner ce travail.

Le travail de Gordon ([14]):

Le but de ce travail était de comparer les cinq meilleures méthodes issues du classement de Milligan et Cooper dans le cas où les ensembles de points contiennent une structure "emboitée".

Le travail de Gordon était donc basé sur l'idée suivante: ne spécifier qu'une valeur pour c (nombre de classes) peut entrainer une représentation "trompeuse" de la structure présente dans l'ensemble de données. Son but était alors de comparer la capacité des cinq meilleures méthodes de Milligan et Cooper à détecter quand plusieurs valeurs de c différentes et bien distinctes (3 et 12 dans son article) pouvaient être appropriées. Ce qui correspondait bien au fait qu'il y ait une structure dans les données à différentes échelles.

Pour tester toutes les méthodes, Gordon a du les modifier afin qu'elles puissent indiquer plus d'une valeur pour c. Les modifications, la manière dont les données ont été générées et les résultats obtenus sont disponibles dans l'article de Gordon mais ne nous intéressent pas ici, le but de notre travail étant tout à fait différent. En effet, le cas examiné par Gordon était très particulier alors que nous traiterons des mêmes méthodes (plus une autre) que lui mais dans un cadre beaucoup plus général. Notons que la méthode qui a été ajoutée est celle appelée CCC (Cubic Clustering Criterion). Cette méthode a été ajoutée car elle possède un intérêt tout particulier. En effet, elle est présente dans SAS, un des logiciels statistiques les plus évolués et utilisés.

Examinons enfin comment nous avons procédé.

Notre travail:

En fait, nous avons refait exactement le même travail que A. Hardy à ceci près: les règles d'arrêt choisies n'étaient plus celles disponibles dans Clustan mais les six meilleures méthodes issues de l'article de Milligan et Cooper[1].

1. Toutes ces règles seront présentées au chapitre 5.

De plus, nous avons testé les méthodes sur certains des ensembles de données de A. Hardy auxquels nous avons ajouté quelques exemples présentant des structures particulières qui auraient pu poser des problèmes (et qui donnaient des résultats intéressants). Remarquons qu'à nouveau, tous ces ensembles de points étaient à deux dimensions et présentaient des classes "naturelles" bien séparées[1]. Ensuite, nous avons appliqué quatre méthodes de classification hiérarchiques[2] à chaque exemple pour obtenir une plus grande variété de résultats. Enfin, pour chaque exemple et à chacune des classifications obtenues, nous avons appliqué les six méthodes de détermination du nombre de classes choisies et nous avons comparé les résultats. A nouveau, les mêmes tableaux que ceux présentés dans le travail de A. Hardy ont été utilisés pour indiquer les résultats.

Cependant, il paraissait plus qu'intéressant de reprendre les résultats obtenus par la méthode et les critères basés sur les hypervolumes pour les inclure dans nos conclusions, ces méthodes surclassant généralement les autres dans le travail de A. Hardy.
Voici enfin un petit plan de travail qui va vous aider à comprendre comment nous avons résolu ce problème :

- le chapitre 3 contient :
  - une partie du travail de A. Hardy.
  - une synthèse des résultats obtenus en appliquant les méthodes basées sur les hypervolumes aux jeux de données ajoutés dans ce mémoire.
- le chapitre 4 contient entre autres:
  - les résultats des méthodes de ce mémoire obtenus avec tous les jeux de données.
  - les conclusions sur ces résultats en considérant aussi les résultats obtenus par les méthodes basées sur les hypervolumes.

1. Comme nous l'avions signalé plus tôt dans le paragraphe sur les classes "naturelles".
2. Celles utilisées par A. Hardy et présentées au chapitre 1.

Chapitre 3

Comparaison entre les méthodes de détermination du nombre de classes disponibles dans Clustan et celles basées sur le critère des hypervolumes

3.1 Introduction

Ce chapitre contient trois parties :

- La première explique les tableaux qui sont utilisés dans ce chapitre et dans le chapitre 4.
- La deuxième contient une partie des résultats du travail de A. Hardy.
- Enfin, la troisième précise les résultats des méthodes basées sur le critère des hypervolumes sur les exemples ajoutés dans ce mémoire et que l'on peut retrouver au chapitre 4.

3.2 Comment lire les tableaux

Rappelons que nous devrons appliquer un certain nombre de méthodes de classification pour chaque jeu de données. Ensuite, à chaque classification obtenue, nous appliquerons les différentes méthodes de détermination du nombre de classes.
Donc, pour chaque ensemble de points, nous retrouverons un tableau de la forme :

"Nom" du jeu de données		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche							
Voisin le plus éloigné							
Moyenne							
Wards							

où

- les différentes lignes du tableau correspondent aux méthodes de classification utilisées.
- la deuxième colonne du tableau indique si les méthodes de classification retrouvent les classes naturelles pour k fixé et correspondant au bon nombre de classes. Il y a un "+" si c'est le cas et un "-" sinon.
- les autres colonnes du tableau indiquent le nombre de classes donné par les différentes méthodes de détermination du nombre de classes utilisées (que l'on note $M1, M2, \ldots, M6$ dans ce chapitre et $M_1, M_2, \ldots, M_6$ dans le chapite 4).

3.3 Le travail de A.Hardy

3.3.1 Méthodes pour la détermination du nombre de classes dans un ensemble de données.

Les trois premières méthodes sont basées sur le critère des hypervolumes. ([17]).

La méthode du coude (M1)

Cette méthode est bien connue. Elle consiste à tracer la courbe du critère de classification W(P,k) en fonction de k, le nombre de classes. Un "coude" dans cette courbe est sensé indiquer le nombre de classes à retenir. Comme le soulignent à juste titre certains auteurs ([13]), cette procédure n'est pas toujours très fiable : de grandes variations peuvent intervenir dans la valeur du critère, même lorsqu'on est en présence d'un ensemble de données sans structure. Nous montrerons cependant que cette méthode, associée au critère des hypervolumes, donne de meilleurs résultats.

Méthode basée sur l'estimation d'un ensemble convexe (M2)

Cette méthode provient du même problème que la méthode des hypervolumes et qui a été proposé par D.G.Kendall : "Etant donné la réalisation d'un processus de Poisson homogène d'intensité inconnue à l'intérieur d'un ensemble convexe compact D, trouver D en utilisant des méthodes d'inférence statistique". Rappelons la solution de ce problème qui fut trouvée par Ripley et Rasson ([30]) et Rasson ([28]).

L'estimation proposée est $D' = g(H(x)) + cs(H(x))$ où

- $x = \{x_1, x_2, ..., x_n\}$ est la réalisation du processus de Poisson dans D;
- H(x) est l'enveloppe convexe de x;
- g(H(x)) est le centroide de H(x);
- $s(H(x)) = H(x) - g(H(x))$.

Nous avions déjà signalé qu'il s'agissait d'une expansion homothétique de l'enveloppe convexe à partir de son centre de gravité. La constante de dilatation c peut être estimée par $c = \sqrt{\frac{n}{n-v_n}}$ où v_n est le nombre de sommets de H(x) ([26]).

De plus, signalons que la réalisation d'un processus de Poisson dans la somme

de k sous-ensembles C_1, C_2, ..., C_k peut être considérée comme la réalisation de k processus de Poisson de même intensité dans les k sous-ensembles C_1, C_2, ..., C_k ([27]).
Notons $C_i'^k$ l'estimation de l'ensemble convexe C_i^k où C_i^k est le ième sous-ensemble de la partition de E en k classes, c'est-à-dire $P \in P_k$ et $P = \{C_1^k, C_2^k, ..., C_k^k\}$.

La règle de décision suivante est alors proposée ([17]) :

- si, pour tout $\{i,j\} \subset \{1,2,...,t\}$, $i \neq j$: $C_i'^t \bigcap C_j'^t = \emptyset$,
  si pour tout $s \in N\backslash\{0,1\}$, $s < t$ et pour tout $\{i,j\} \subset \{1, 2, ..., s\}$, $i \neq j$: $C_i'^s \bigcap C_j'^s = \emptyset$, alors on conclut que la partition naturelle contient au moins t classes et on examine la partition optimale en (t+1) classes;
- s'il existe $\{i,j\} \subset \{1,2,...,t\}$, $i \neq j$: $C_i'^t \bigcap C_j'^t \neq \emptyset$,
  si pour tout $s \in N\backslash\{0,1\}, s < t$, et pour tout $\{i,j\} \subset \{1,2,...,s\}$:
  $C_i'^s \bigcap C_j'^s = \emptyset$, alors on conclut que l'ensemble de données contient t-1 classes naturelles;
- si $C_1'^2 \cap C_2'^2 \neq \emptyset$, alors on conclut qu'on est en présence d'un ensemble de données sans structure.

Un test du quotient de vraisemblance (M3)

Comme il existe un modèle explicite associé à la méthode des hypervolumes ([17]), on peut formuler un test du quotient de vraisemblance pour la présence d'une structure dans un ensemble de données ([17]).
Soit X_1, X_2, ,..., X_n la réalisation d'un processus de Poisson homogène dans t ensembles convexes d'un espace euclidien à m dimensions.

Nous testons si une subdivision en k classes est significativement meilleure qu'une subdivision en k-1 classes, i.e. H_0: t=k contre H_1: t=k-1.

Notons

- $C = \{C_1, C_2, ..., C_k\}$ la partition optimale en k classes ;
- $D = \{D_1, D_2, ..., D_{k-1}\}$ la partition optimale en k-1 classes.

Le quotient de vraisemblance prend la forme ([17]) :

$$Q(x) = \left(\frac{W(P,k)}{W(P,k-1)}\right)^n.$$

La région critique prendra la forme :

$$RC = \{S > C\} = \left\{S = \frac{W(P,k)}{W(P,k-1)} > C\right\}.$$

Malheureusement, la distribution de la statistique S n'est pas connue (comme c'est souvent le cas pour les statistiques de test associées à d'autres méthodes de classification). Mais S possède une propriété intéressante. On a en effet $S(x) \in [0, 1]$.

Nous pouvons donc en pratique utiliser la règle de décision suivante: rejeter H_0 si S prend de grandes valeurs, c'est-à-dire si S est proche de 1. Si k_0 est la première valeur pour laquelle on rejette H_0, on considèrera $k_0 - 1$ comme le nombre approprié de classes naturelles.

3.4 Autres méthodes

Les quatre autres méthodes sont bien connues dans la littérature relative à la classification. Elles sont disponibles dans le logiciel CLUSTAN ([33]): test de Wolfe (M4), règle du quantile supérieur (M5), règle du contrôle de la moyenne mobile (M6) et test de Marriot (M7).

3.3.2 Résultats

Pour pouvoir comparer les méthodes de détermination du nombre de classes, ont été choisis : six procédures de classification bien connues (saut minimum, saut maximum, centroide, Ward, k-means et hypervolume), six ensembles de données artificielles dont la structure est connue (classes bien séparées, données sans structure, classes allongées, classes non convexes, classes non séparables par un hyperplan, données de Ruspini) et bien entendu les sept méthodes pour la détermination du nombre de classes. Les résultats sont repris dans les tableaux qui suivent.

Premier ensemble de données : classes bien séparées

Cet ensemble de données provient d'une simulation d'un processus de Poisson homogène dans trois classes bien séparées (Figure 1).

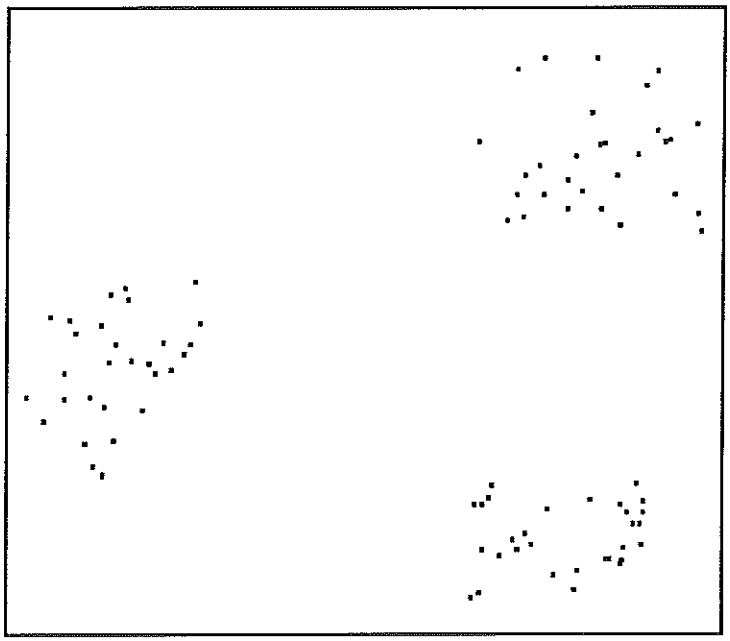


Figure 1 : Trois classes bien séparées.

Classes bien séparées		M1	M2	M3	M4	M5	M6	M7
saut minimum	+	3	–	–	3	2	3	3
saut maximum	+	5	–	–	3	2	3	3
centroide	+	3	–	–	3	3	3	3
Ward	+	3	–	–	3	3	3	3
K-means	+	3	–	–	3	–	–	3
hypervolume	+	3	3	3	–	–	–	3

Table 1 : Trois classes bien séparées.

Ici nous avons un des exemples les plus simples. En effet, les six procédures de classification retrouvent la structure naturelle des données. C'est pourquoi il y a un "+" dans la seconde colonne de la Table 1. Les sept autres colonnes montrent le résultat donné par les sept méthodes pour la détermination du nombre de classes.

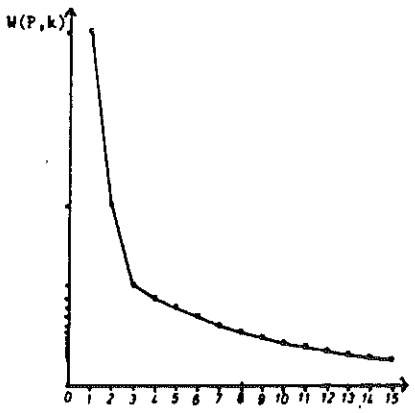


Figure 2 : Critère des hypervolumes.

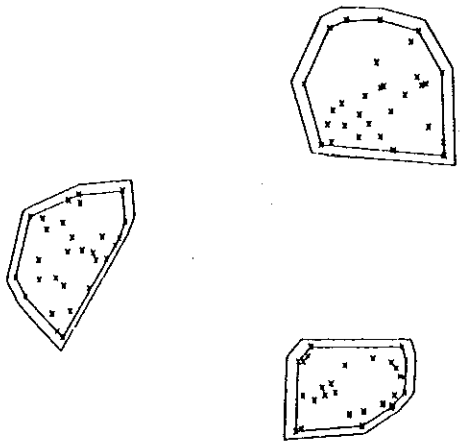


Figure 3a.

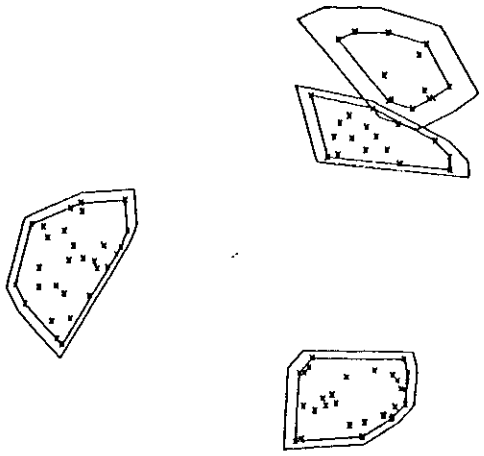


Figure 3b.

Par exemple l'application de la méthode basée sur l'estimation d'un ensemble convexe (M2) et du test du quotient de vraisemblance (M3) aux classifications données par la procédure des hypervolumes nous permet de conclure qu'il y a trois classes dans l'ensemble de données de la Figure 1.

La Figure 2 montre la courbe de décroissance du critère des hypervolumes en fonction du nombre de classes (M1). Un coude apparaît en k=3.
La Figure 3 illustre la méthode M2. Les enveloppes convexes dilatées correspondant aux partitions optimales en 2 classes (respectivement en 3 classes) ne se coupent pas. Par contre deux des expansions homothétiques relatives à la partition en 4 classes présentent une intersection non vide. De plus, la Table 2 montre les différentes valeurs prises par la statistique S associée à la méthode M3. Lorsque k=4, on peut considérer que V(k) est proche de 1. On en conclut que l'ensemble de données de la Figure 1 contient 3 classes naturelles.

k	S(k)
1	-
2	0.51
3	0.56
4	0.86
5	0.89
6	0.88
7	0.87
8	0.89
9	0.90
10	0.88
11	0.91
12	0.89
13	0.90
14	0.91
15	0.90

Table 2 : Valeurs de la statistique S associée à la méthode M3.

Dans ce premier exemple, la méthode du coude donne le résultat attendu avec la plupart des méthodes de classification; les classes sont évidemment très bien séparées. La méthode du contrôle de la moyenne mobile (M6) et le test de Wolfe (M4) donnent aussi le nombre approprié de classes. Remarquons que le test de Marriot (M7) fournit aussi les résultats attendus : la courbe $k^2 det(W)$ atteint son minimum pour $k^* = 3$.

Le signe "–" dans une des sept dernières colonnes des tables de résultats signifie que les hypothèses pour appliquer la méthode ne sont pas satisfaites. Par exemple les méthodes M2 et M3 ne sont valables que si elles sont associées aux classifications obtenues par la méthode des hypervolumes. D'autre part, les méthodes M5 et M6 ne peuvent être appliquées qu'aux méthodes hiérarchiques. Les résultats de la Table 1 montrent cependant que nous devons être très prudents. Même en présence d'un exemple aussi simple que celui présenté ici, toutes les méthodes pour la détermination du nombre de classes ne donnent pas le résultat espéré.

Deuxième ensemble de données : données sans structure

Pour cet exemple, un processus de Poisson a été simulé dans un seul domaine du plan. Les 150 points sont donc distribués indépendamment et uniformément dans ce domaine (Figure 4). En fait, cela permet de tester l'absence de structure d'un ensemble de données ([4]).

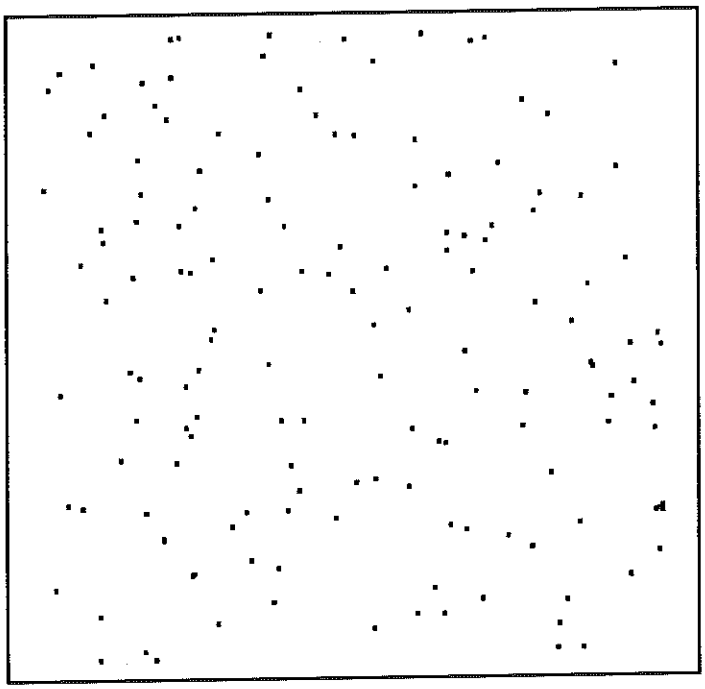


Figure 4 : Données sans structure.

Données sans structure	M1	M2	M3	M4	M5	M6	M7
saut minimum	1	–	–	1	2	2	X
saut maximum	*	–	–	2	2	4	X
centroide	*	–	–	2	3	3	X
Ward	4	–	–	3	2	4	X
K-means	4	–	–	2	–	–	X
hypervolume	1	1	1	–	–	–	X

Table 3 : Données sans structure.

Les résultats de la Table 3 montrent que les trois méthodes basées sur le critère des hypervolumes sont très efficaces lorsqu'il s'agit de tester la présence d'une structure dans un ensemble de données.
Si l'on considère la méthode basée sur l'estimation d'un ensemble convexe

(Figure 6), on remarque que les expansions homothétiques correpondant à la partition optimale en deux classes se coupent. D'autre part la Table 4 reprend les valeurs de la statistique du test de la méthode M3. S(2) = 0.90 est proche de 1. De son côté, M_1 indique que le critère des hypervolumes décroît lentement et uniformément lorsque le nombre de classes augmente (Figure 3).
Dans les trois cas, on en déduit que l'ensemble de données de la Figure 4 ne présente aucune structure particulière. On est donc en présence d'un seul groupe homogène.

Les autres méthodes ne se comportent pas très bien, sauf M1 et M4 lorqu'elles sont associées aux résultats donnés par la procédure du saut minimum. Le test de Marriot n'est pas applicable ici. D'autre part, la présence d'un "*" dans une case signifie que les résultats ne sont pas suffisamment clairs que pour pouvoir conclure.

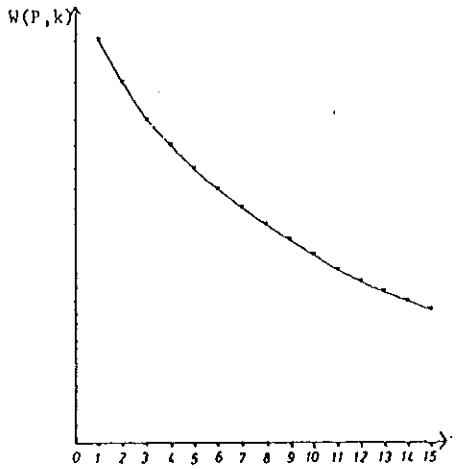


Figure 5 : Méthode du coude.

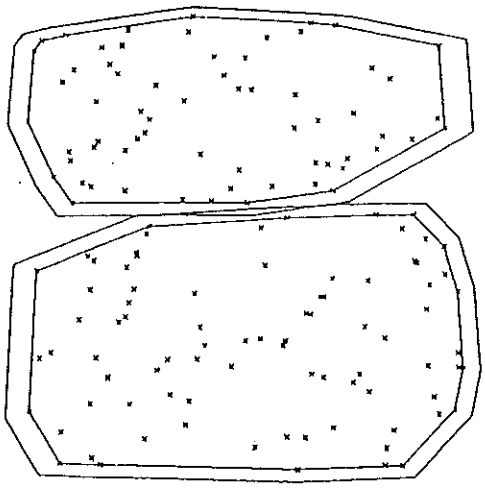


Figure 6 : Méthode basée sur l'estimation d'un ensemble convexe.

k	S(k)
1	-
2	0.90
3	0.90
4	0.92
5	0.92
6	0.93
7	0.92
8	0.93
9	0.93
10	0.93
11	0.92
12	0.94
13	0.94
14	0.93
15	0.94

Table 4 : Valeurs de la statistique S pour la méthode M3.

Troisième ensemble de données: classes non séparables par un hyperplan.

Les données de la Figure 9 contiennent trois classes allongées. Aucune d'elles n'est séparable des autres par un hyperplan.

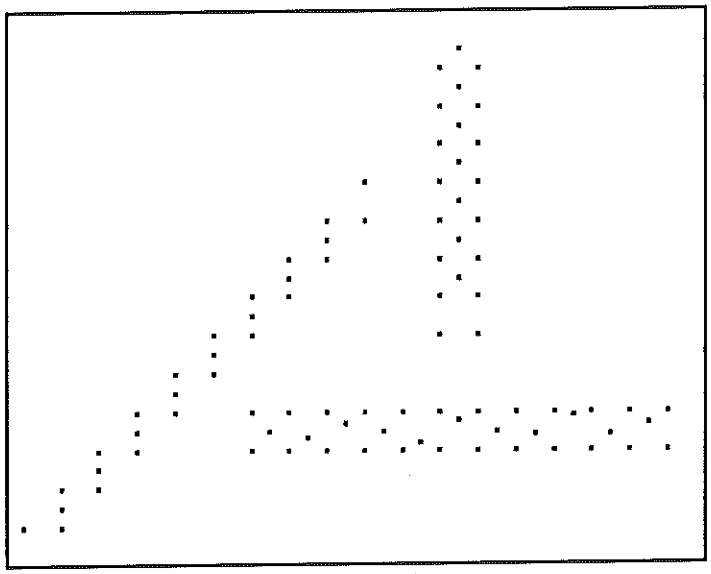


Figure 9 : Classes non séparables par un hyperplan.

Classes non séparables...		M1	M2	M3	M4	M5	M6	M7
saut minimum	+	3	–	–	3	2	3	10
saut maximum	-	*	–	–	3	2	2	10
centroide	-	*	–	–	3	2	2	10
Ward	-	3	–	–	3	2	3	10
K-means	-	3	–	–	3	–	–	10
hypervolume	+	3	*	3	–	–	–	10

Table 6 : Classes non séparables par un hyperplan.

Seules les méthodes du saut minimum et des hypervolumes retrouvent la structure naturelle en trois classes. La méthode du coude, le test de Wolfe et la méthode du contrôle de la moyenne mobile associées à la procédure du saut minimum, ainsi que la méthode du coude et le test du quotient de vraisemblance appliqués à la classification produite par la méthode des hypervolumes, donnent le résultat attendu.

Quatrième ensemble de données: données de Ruspini

Les données de Ruspini sont souvent utilisées pour tester de nouvelles méthodes en classification automatique ([8];[9];[16]; ...). Elles sont composées de 75 points dans le plan. On y reconnaît généralement quatre classes.

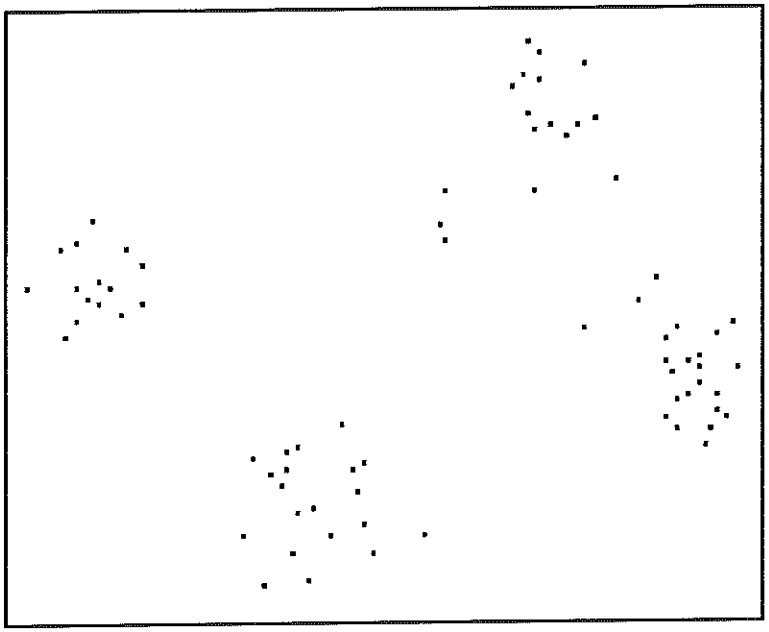


Figure 11 : données de Ruspini.

Données de Ruspini		M1	M2	M3	M4	M5	M6	M7
saut minimum	+	4	–	–	4	2	4	5
saut maximum	+	4	–	–	4	2	4	5
centroide	+	4	–	–	4	2	4	5
Ward	+	4	–	–	4	2	4	5
K-means	+	4	–	–	4	–	–	5
hypervolume	+	4	4	4	–	–	–	5

Table 8 : Données de Ruspini.

Si l'on fixe le nombre de classes à quatre, les six procédures de classification retrouvent la structure naturelle des données de Ruspini. Toutes les méthodes pour la détermination du nombre de classes ne donnent cependant pas le résultat attendu. La Figure 12 illustre la méthode M2 et la Table 8 reprend les valeurs de la statistique S (méthode M3). On en déduit que la

partition naturelle des données contient quatre classes.

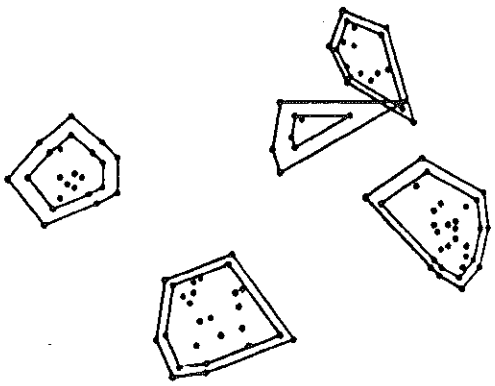


Figure 12 : Méthode M2 et données de Ruspini.

k	S(k)
1	-
2	0.49
3	0.74
4	0.70
5	0.84
6	0.86
7	0.85
8	0.87
9	0.88
10	0.85
11	0.87
12	0.88
13	0.89
14	0.88
15	0.87

Table 8 : Méthode M3 et données de Ruspini.

3.3.3 Conclusions

Voici maintenant une partie des conclusions tirées par A. Hardy dans son article suite aux résultats précédents.

Le premier problème important pour découvrir la structure d'un ensemble de données est le choix d'une "bonne" procédure de classification. En effet, certaines méthodes pour la détermination du nombre de classes donnent le nombre attendu de classes alors que les classes trouvées par la méthode de classification ne sont pas les classes naturelles. Il est donc recommandé de tenir compte, dans une analyse de classification, des hypothèses liées aux méthodes, et des biais qui leur sont associés (effet de chaînage, tendance à favoriser des classes sphériques ou de même taille, difficulté à prendre en compte la présence de points isolés, etc.).

Dans l'ensemble, les trois méthodes basées sur le critère des hypervolumes donnent de bons résultats.

La méthode du coude (M1) est en général très subjective. Il n'est pas toujours facile de voir où se trouve le coude qui indique le nombre de classes présentes. Néanmoins, la méthode du coude associée au critère des hypervolumes conduit souvent à des résultats clairs.

La méthode basée sur l'estimation d'un ensemble convexe (M2) donne très souvent les résultats attendus. Elle rencontre quand même certaines difficultés quand les classes sont allongées et contiennent très peu de points. A. Hardy indique alors qu'il faudrait certainement chercher une valeur plus adéquate pour le coefficient de dilatation de l'enveloppe convexe lorsqu'on traite le problème de la détermination du nombre de classes.

Le test du quotient de vraisemblance (M3) semble la méthode la plus intéressante. Bien que l'on ne connaisse pas la distribution de la statistique du test, les conclusions sur le nombre de classes sont claires et pertinentes. Ceci est dû à la nature du critère mais aussi au fait que cette statistique est toujours comprise entre 0 et 1.

En analysant les résultats obtenus sur les exemples présentés ici (et aussi sur une beaucoup d'autres exemples), A. Hardy recommande, losrqu'on est en présence de données réelles, d'appliquer plusieurs procédures de classification et plusieurs méthodes pour la détermination du nombre de classes. La comparaison des résultats obtenus devrait alors permettre d'avoir des informations sur la structure (si il y en a une) des données et sur les classes (taille, forme, orientation, séparation, présence de points isolés, ...). C'est sur base de cette information qu'il reste à retenir la méthode de classification la plus adéquate et la meilleure partition des données.

3.4 Résultats supplémentaires

En ce qui concerne les exemples qui seront ajoutés au chapitre 4, c'est-à-dire :

- Un ensemble de données avec deux classes bien séparées.
- Un ensemble de données avec deux classes parallèles.
- Un ensemble de données avec trois classes parallèles.
- Un ensemble de données où les classes ne sont pas convexes (données en sourire).
- Un ensemble de données présentant deux classes dont une a beaucoup plus de points que l'autre.

En fait, seul l'ensemble de données en sourire pose problème. En effet, l'hypothèse de convexité des classes n'est pas respectée, et en plus les classes naturelles ne sont pas séparables par un hyperplan. On constate alors que la méthode des hypervolumes ne retrouve pas les classes naturelles.

Signalons aussi que des ensembles de données avec deux classes non convexes et avec deux classes parallèles un peu différents de ceux du chapitre 4 avaient été traités dans l'article de A. Hardy.

Chapitre 4

Comparaison des six meilleures méthodes de détermination du nombre de classes de l'article de Milligan et Cooper avec celles basées sur le critère des hypervolumes

4.1 Introduction

Cette introduction afin de signaler qu'à part une fois dans la partie concernant la méthode du "C-index", les notations des articles de base ont été gardées dans ce mémoire. Toute personne désireuse d'en savoir plus ou de retravailler ces diverses méthodes pouvant ainsi s'y retrouver. Notons quand-même que deux autres petites choses ont été changées : le nombre d'objets et le nombre de classes ont toujours été notés respectivement n et k, même si ce n'était pas le cas dans l'article de référence.

4.2 Présentation du programme utilisé

Signalons pour commencer que la version initiale de ce programme a été réalisée par A.D. Gordon.
Maintenant, nous allons expliquer un peu à quoi servait ce programme et comment il procédait.
En fait, le programme calculait[1] pour chaque méthode de classification[2], la hiérarchie de partitions correspondante. De plus, pendant la formation de chaque hiérarchie de partitions, il calculait aussi les valeurs des six indices (pour la détermination du nombre de classes) pour chaque partition de la hiérarchie. Enfin, l'algorithme utilisé procédait de façon agglomérative. Il partait donc de la partition en n classes pour terminer par la partition en une classe.
Aussi, l'indice CCC n'était pas calculé par le programme. En effet, nous avons trouvé ses valeurs grâce au logiciel SAS. Mais celui-ci procédait de façon tout à fait analogue au progamme utilisé dans ce mémoire.
On obtenait donc à chaque fois un résultat de ce genre:

Ensemble de données 1.
Méthode du voisin le plus proche.

	M_1	M_2	M_3	M_4	M_5	M_6
k = n						
n-1						
⋮						
1						

où les colonnes du tableau contenaient les valeurs des indices correspondant aux méthodes $M_1, \ldots, M_6$, pour chaque partition en $n, n-1, \ldots, 1$ classes.

Méthode du voisin le plus éloigné.
⋮

Ensuite, pour chaque tableau, il fallait appliquer la règle de décision du nombre de classes correspondant à chaque méthode $M_1, \ldots, M_6$ suivant les différentes valeurs des indices. Par exemple, certaines méthodes exigent de prendre la valeur maximale de l'indice parmi toutes celles correspondant aux partition en $n, n-1, \ldots, 1$ classes.

1. A chaque fois qu'il était exécuté pour un certain jeu de données.
2. Rappelons que les méthodes utilisées étaient la méthode du voisin le plus proche, la méthode du voisin le plus éloigné, la méthode de Ward et la méthode de la moyenne.

4.3 Présentation des six méthodes de l'article de Milligan et Cooper

4.3.1 La méthode Gamma M_1 ([20])

Description

Soit un ensemble de n objets : $\{o_1, o_2, \ldots, o_n\}$.

Notons : $d(o_i, o_j) \equiv$ distance euclidienne entre o_i et o_j.
Supposons que ces objets sont partitionnés en k classes.

Définissons

$$T_l(o_i, o_j) = \begin{cases} 0 & \text{si } o_i \text{ et } o_j \text{ sont dans la même classe} \\ 1 & \text{sinon} \end{cases}$$

De plus, si $T_l(o_i, o_j)$ vaut 0 (o_i et o_j sont dans la même classe), définissons aussi:

$$n_l(o_i, o_j) = \sharp\{\{o_r, o_t\} : T_l(o_r, o_t) = 1 \ et \ d(o_r, o_t) < d(o_i, o_j)\}$$

En fait, $n_l(o_i, o_j)$ représente le nombre de couples d'objets n'appartenant pas à la même classe et qui sont plus proches que o_i et o_j.

Nous pouvons maintenant définir l'indice α_l :

$$\alpha_l = \frac{\sum_{i<j} n_l(o_i, o_j)}{\max \sum_{i<j} n_l(o_i, o_j)}$$

où

- le maximum est pris sur toutes les partitions possibles en k classes en gardant le même nombre d'objets par classe.
- les sommes sont sur les paires d'objets $\{o_i, o_j\}$ qui appartiennent à la même classe (telles que $T_l(o_i, o_j) = 0$).

Or, $\sum_{i<j} n_l(o_i, o_j)$ représente la somme sur les paires d'objets appartenant à une même classe, du nombre de paires d'objets n'appartenant pas à une même classe et qui sont plus proches que la paire d'objets considérée dans la somme.

Donc, α_l **est la proportion de paires d'objets qui sont dans des "mauvais groupes".**
Et cela dans le sens où deux objets d'une même classe sont strictement plus proches que deux objets de classes différentes (cela correspond bien à nos définitions des premiers chapitres).

Enfin, l'indice γ est :

$$\boxed{\gamma = 1 - 2\alpha_l}$$

Remarquons que l'indice γ varie de -1 à $+1$.

En effet: Si aucun objet n'est mal placé, alors $\sum_{i<j} n_l(o_i, o_j) = 0$
Par conséquent, $\alpha_l = 0$ et $\gamma = 1$.
Si tous les objets sont mal placés, alors $\alpha_l = 1$
et $\gamma = -1$.

Cela montre aussi que γ vaut 1 si et seulement si la partition est "parfaite" (c'est-à-dire que, par rapport à notre définition, tous les objets sont "bien placés").

Bien évidement, **la méthode Gamma cherche la valeur maximale de l'indice, celle-ci devant être le plus proche possible de 1.**
Rappelons que dans notre cas, cette valeur était cherchée parmi toutes celles correspondant aux partitions de la hiérarchie de partitions obtenue par la méthode de classification utilisée.

Que vaut $\max \sum_{i<j} n_l(o_i, o_j)$ **?**

Soit : x, le nombre de couples d'objets appartenant à une même classe.
y, le nombre de couples d'objets appartenant à des classes différentes.

Combien peut-il y avoir au maximum d'objets "mal placés" ?
On doit comparer chaque paire d'objets appartenant à une même classe avec chaque paire d'objets appartenant à des classes différentes. Donc, on fait $x * y$

comparaisons. Le nombre maximum de cas "défavorables" est alors de $x * y$. **Mais**, de ce nombre, il faut retrancher tous les cas où il y a égalité, c'est-à-dire où des objets appartenant à la même classe sont à la même distance que des objets appartenant à des classes différentes. En effet, ces objets ne sont pas considérés comme mal placés car, dans la définition de $n_l(o_i, o_j)$, on a $d(o_r, o_t) < d(o_i, o_j)$.

En résumé:

$$\max \sum_{i<j} n_l(o_i, o_j) \equiv x\ y\ -\ \text{"}\textit{égalités}\text{"}.$$

4.3.2 La méthode de Duda et Hart M_2 ([10])

Description

Définissons

- $\mathcal{X}_i \equiv$ l'ensemble des points du i-ème groupe.
- $m_i \equiv$ la moyenne des points du i-ème groupe.
- $J(k) = \sum_{i=1}^{k} \sum_{x \in \mathcal{X}_i} \|x - m_i\|^2$.

Il est évident que $J(k)$ diminue de façon monotone avec k car la somme des carrés des erreurs peut être réduite à chaque fois que k augmente en formant un nouveau groupe avec un seul objet. On peut aussi montrer que si il y a k_0 groupes bien séparés, on s'attend à ce que $J(k)$ diminue rapidement jusque $k = k_0$ et diminue beaucoup moins rapidement ensuite jusqu'à ce qu'il atteigne 0 quand k=n.

Maintenant, on voudrait pouvoir voir, grâce à une amélioration statistiquement significative de $J(k)$ que décrire l'ensemble de points avec $k+1$ groupes est mieux adapté qu'avec k groupes.
Une façon formelle de faire est d'avancer l'hypothèse nulle qu'il y a exactement k groupes et de calculer la distribution d'échantillonage pour $J(k+1)$ sous cette hypothèse. Cette distribution nous dirait alors à quelle genre d'amélioration on doit s'attendre pour $J(k)$ quand une description de l'ensemble de points en k groupes est correcte. La procédure de décision serait alors d'accepter l'hypothèse nulle si la valeur de $J(k+1)$ tombe au-delà d'une limite correspondant à une probabilité raisonnable de "rejet à tort" de l'hypothèse nulle.
Malheureusement, on ne peut généralement rien faire d'autre que d'estimer grossièrement la distribution de $J(k+1)$. Les solutions obtenues ne sont alors pas au-dessus de tout soupçon et le problème statistique de tester la validité des groupes est encore irrésolu.

Duda et Hart ont alors essayé l'approximation ci-dessous du critère de la "somme des carrés des erreurs".
Supposons que l'on ait un ensemble $\mathcal{X}$ de n points et que nous voulions décider si oui ou non on peut justifier l'hypothèse qu'ils forment plus qu'un groupe.

Avançons l'hypothèse nulle:

H_0 : les points viennent d'une population normale de moyenne μ et de matrice de covariance $\sigma^2 I$.

Si cette hypothèse était vraie, tout groupe trouvé parmi les points aurait été formé par chance, et toute diminution observée de la somme des carrés des erreurs dans la classification n'aurait aucune signification.
Considérons la somme des carrés des erreurs $J_e(1)$ comme une variable aléatoire, car elle dépend de l'ensemble particulier de points choisit.

$$J_e(1) = \sum_{x \in \mathcal{X}} \|x - m\|^2$$

où m est la moyenne des n objets.

Duda et Hart affirment alors que sous l'hypothèse nulle, la distribution de $J_e(1)$ est approximativement normale, de moyenne $nd\sigma^2$ [1] et de variance $2nd\sigma^4$.

Ils supposent ensuite que l'on peut diviser l'ensemble des objets en deux sous-ensembles $\mathcal{X}_1$ et $\mathcal{X}_2$ de façon à minimiser $J_e(2)$ où:

$$J_e(2) = \sum_{i=1}^{2} \sum_{x \in \mathcal{X}_i} ||x - m_i||^2$$

où m_i est la moyenne des objets de $\mathcal{X}_i$.

Ils expliquent alors que sous l'hypothèse nulle, cette partition est n'est pas la meilleure mais il en résulte néanmoins une valeur de $J_e(2)$ plus petite que $J_e(1)$. Si on connaissait la distribution d'échantillonage de $J_e(2)$, on pourrait alors déterminer à quel point $J_e(2)$ devrait être petit pour que l'on soit obligé d'abandonner l'hypothèse nulle d'un seul groupe. Comme on n'a pas la "solution analytique" de la partition optimale, on ne peut pas trouver une solution exacte pour la distribution d'échantillonage. Heureusement, Duda et Hart disent que l'on peut obtenir une bonne estimation en considérant la partition "sous-optimale" obtenue en prenant un hyperplan passant par la moyenne de l'échantillon.
Pour n grand, on peut montrer que la somme des carrés des erreurs pour cette partition est approximativement normale, de moyenne $n(d - \frac{2}{\pi})\sigma^2$ et de variance $2n(d - \frac{8}{\pi^2})\sigma^4$. Ce résultat s'accorde bien avec le fait que $J_e(2)$ est plus petit que $J_e(1)$, puisque la moyenne pour $J_e(2)$ -$n(d - \frac{2}{\pi})\sigma^2$- est plus petite que la moyenne pour $J_e(1)$ -$nd\sigma^2$-. Mais pour pouvoir être considérée comme significative, la réduction de la somme des carrés des erreurs doit être plus grande que cela. Duda et Hart expliquent à ce moment que l'on peut

1. Où d est le nombre de dimensions des données.

obtenir une approximation de la valeur critique pour $J_e(2)$ en supposant que la valeur sous-optimale est presque optimale, en utilisant l'approximation normale pour la distribution d'échantillonage et en estimant σ^2 par :

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{x \in \mathcal{X}} \|x - m\|^2 = \frac{1}{nd} J_e(1)$$

Le résultat final est alors :
Rejeter l'hypothèse nulle "au niveau p" si

$$\frac{J_e(2)}{J_e(1)} < 1 - \frac{2}{nd} - \alpha \sqrt{\frac{2(1 - \frac{8}{\pi^2 d})}{nd}}$$

où α est déterminé par

$$p = 100 \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2u^2}} du.$$

Dans notre cas, on va inverser la procédure. Rappelons-nous en effet que, à chaque étape, les méthodes hiérarchiques agglomératives regroupent les deux classes les plus proches. On utilise alors le test pour décider si oui ou non la fusion de deux groupes est justifiée.
Il est évident que ce test s'applique très bien aux hiérarchies de partitions obtenues par des méthodes hiérarchiques. Nous avons alors qu'à chaque étape, le test n'est appliqué qu'aux points des classes concernées par la fusion (ce sera le cas aussi pour la méthode de Beale).

En pratique, la méthode calcule à chaque étape (pour toute partition en k classes, en partant de $k = n$) :

$$\frac{-\frac{J_e(2)}{J_e(1)} + 1 - \frac{2}{\pi d}}{\sqrt{\frac{2(1 - \frac{8}{\pi^2 d})}{nd}}} \quad .$$

On a alors que **dès que cette expression dépasse la valeur choisie pour α (pour $k = k_0$), on rejette l'hypothèse nulle qui correspond à l'existence d'un seul groupe. C'est donc que la fusion des deux groupes concernés n'était plus justifiée[2]. On prend alors comme bon nombre de classes $k_0 + 1$.**

Dans la littérature, plusieurs valeurs de α ont été proposées. En effet, Milligan et Cooper ont indiqué que 3.20 semblait être la valeur qui donnait les résultats optimaux ([25]). Tandis que Gordon utilisait quant à lui 4 comme valeur pour α ([14]) tout en signalant que devoir spécifier ainsi une valeur critique était un inconvénient de la méthode.

2. Chaque α correspond à un niveau p. En général, dans la pratique, on choisit alors α pour avoir les meilleurs résultats possibles.

4.3.3 La méthode de Beale M_3 ([3])

Notons :

- W_1, la somme des carrés des distances à l'intérieur d'un ensemble de points.
- W_2, la somme des carrés des distances à l'intérieur des deux groupes obtenus en divisant l'ensemble initial en deux.

Le test proposé par Beale permet, tout comme celui de Duda et Hart, de voir si la fusion de deux groupes de points est justifiée ou pas.

Ce test implique la comparaison de

$$\frac{(\frac{W_1-W_2}{W_2})}{((\frac{n-1}{n-2})2^{\frac{2}{p}}-1)}$$

avec une distribution F de Fisher-Snedecor à p degrés de liberté au numérateur et (n-2)p degrés de liberté au dénominateur.

Tout comme pour le test de Duda et Hart, si k_0 est la première valeur qui conduit à un rejet de l'hypothèse correspondant à la fusion de deux groupes de points, le test de Beale indique que le bon nombre de classes est $k_0 + 1$.

Dans notre cas, p vaut 2 et $(n-2)p$ est toujours plus grand que 120. Ce qui conduit à des valeurs de $F_{p,(n-2)p}$ de 5.30 pour un niveau de précision 0.005 et de 4.61 pour un niveau de précision de 0.01. Ces niveaux donneront toujours les mêmes résultats, sauf pour les données de Ruspini (voir les résultats concernant cet exemple pour plus de précisions).

Remarque : au départ, un test plus général permettait de tester l'existence de $C1$ groupes contre l'existence de $C2$ groupes.

4.3.4 La méthode de Calinski et Harabasz M_4 ([5])

Notations

Supposons que nous ayons n objets classés en k groupes. Dans cette section, nous noterons:

- $d_{ij} \equiv$ distance euclidienne entre les objets i et j.
- $\bar{d}^2 = \sum_{i=1}^{n} \sum_{j=1 j\neq i}^{n} \frac{d_{ij}^2}{\frac{n(n-1)}{2}} \equiv$ moyenne des d_{ij}^2.
- $\bar{d}_g^2 = \sum_{i=1}^{n_g} \sum_{j=1 j\neq i}^{n_g} \frac{d_{ij}^2}{\frac{n_g(n_g-1)}{2}} \equiv$ moyenne des d_{ij}^2 dans le g-ième groupe.
- R=B+W avec
  - $R \equiv$ matrice de dispersion totale.
  - $B \equiv$ matrice de dispersion inter-groupes.
  - $W \equiv$ matrice de dispersion intra-groupes.

  Les éléments des matrices R, W et B étant les suivants:
  - $r_{jl} = \sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)$
  - $w_{jl} = \sum_{c=1}^{k}\sum_{i=1}^{n_c}(x_{ij} - \bar{x}_{jc})(x_{il} - \bar{x}_{lc})$
  - $b_{jl} = \sum_{c=1}^{k} n_c(\bar{x}_{jc} - \bar{x}_c)(\bar{x}_{lc} - \bar{x}_c)$

  avec
  - k: nombre de groupes.
  - n_c: effectif du groupe C.
  - N: effectif de la population.
  - p: nombre de variables.
  - j, l: indices de variables, $1 \leq j \leq p$, $1 \leq l \leq p$.
  - $\bar{x}_j$: moyenne générale de la variable j.
  - $\bar{x}_{lc}$: moyenne de la variable l dans le groupe C.
- $BGSS \equiv$ Between Group Sum of Squares.

- $WGSS \equiv$ Within Group Sum of Squares.

- $TSS \equiv$ Total Sum of Squares.

Nous avons alors :

- Trace$W$=$WGSS$=TraceR_1+...+TraceR_k,

$$\text{où Trace}R_g = n_g^{-1}\left(d^2_{12(g)} + d^2_{13(g)} + \ldots + d^2_{n_{g-1},n_g(g)}\right).$$

- Trace$R$=$\frac{1}{n}(d^2_{12} + d^2_{13} + \ldots + d^2_{n-1,n})$.

Description

Le critère du rapport de variance (VRC) est:

$$VRC = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}}$$

Bien qu'il n'y ait aucune justification théorique en probabilité pour utiliser VRC, ce critère possède quelques propriétés mathématiques intéressantes.

1. $TSS = \frac{1}{2}(n-1)\bar{d}^2$.

   En effet,
   $\frac{1}{2}(n-1)\bar{d}^2 = \frac{\frac{1}{2}(n-1)(d^2_{12}+d^2_{13}+\ldots+d^2_{n-1,n})}{\frac{n(n-1)}{2}} = \frac{1}{n}(d^2_{12} + d^2_{13} + \ldots + d^2_{n-1,n})$.

2. $WGSS = \frac{1}{2}\left((n_1-1)\bar{d}_1^2 + \ldots + (n_k-1)\bar{d}_k^2\right)$.

   En effet,
   comme Trace$W$=TraceR_1+...+TraceR_k, il suffit d'appliquer
   le résultat précédent pour chaque TraceR_i ($\equiv$ TSS du groupe i).

3. $BGSS = \frac{1}{2}\left((k-1)\bar{d}^2 + (n-k)A_k\right)$
   où $A_k = \frac{1}{n-k}\left((n_1-1)(\bar{d}^2 - \bar{d}_1^2) + \ldots + (n_k-1)(\bar{d}^2 - \bar{d}_k^2)\right)$.

   En effet:

   - R=B+W .
     Par conséquent, Trace$B$=Trace$R$-Trace$W$
     et Trace$B$=$\frac{1}{2}(n-1)\bar{d}^2$-$\left[\frac{1}{2}\left((n_1-1)\bar{d}_1^2 + \ldots + (n_k-1)(\bar{d}^2 - \bar{d}_k^2)\right)\right]$.

- Or:

$$\begin{aligned}
&\tfrac{1}{2}\left((k-1)\bar{d}^2+(n-k)A_k\right)\\
&=\tfrac{1}{2}\left[(k-1)\bar{d}^2+\tfrac{n-k}{n-k}\left((n_1-1)(\bar{d}^2-\bar{d}_1^2)+\ldots+(n_k-1)(\bar{d}^2-\bar{d}_k^2)\right)\right]\\
&=\tfrac{1}{2}\left[\bar{d}^2(n-1)-\bar{d}_1^2(n_1-1)-\ldots-\bar{d}_k^2(n_k-1)\right]\\
&\quad \text{car } (k-1)+(n_1-1)+\ldots+(n_k-1)=(n-1).
\end{aligned}$$

Maintenant, nous pouvons écrire:

$$\boxed{VRC=\frac{\left(\bar{d}^2+\frac{n-k}{k-1}A_k\right)}{(\bar{d}^2-A_k)}}$$

En effet:

- $VRC=\frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}}$.

  Mais,

  - $\frac{BGSS}{k-1}=\frac{\frac{1}{2}\left((k-1)\bar{d}^2+(n-k)A_k\right)}{k-1}=\frac{1}{2}\left(\bar{d}^2+\frac{n-k}{k-1}A_k\right)$.
  - $\frac{WGSS}{n-k}=\frac{\frac{1}{2}\left((n_1-1)\bar{d}_1^2+\ldots+(n_k-1)\bar{d}_k^2\right)}{n-k}$.

    Or,

$$\begin{aligned}
(\bar{d}^2-A_k)&=\bar{d}^2-\tfrac{1}{n-k}\left((n_1-1)(\bar{d}^2-\bar{d}_1^2)+\ldots+(n_k-1)(\bar{d}^2-\bar{d}_k^2)\right)\\
&=\bar{d}^2-\tfrac{1}{n-k}\bar{d}^2\left((n_1-1)+\ldots+(n_k-1)\right)\\
&\qquad+\tfrac{1}{n-k}\left((n_1-1)\bar{d}_1^2+\ldots+(n_k-1)\bar{d}_k^2\right)\\
&=\tfrac{1}{n-k}\left((n_1-1)\bar{d}_1^2+\ldots+(n_k-1)\bar{d}_k^2\right)
\end{aligned}$$

    et donc: $\frac{WGSS}{n-k}=\frac{1}{2}(\bar{d}^2-A_k)$.

- En remplaçant, nous obtenons:

$$VRC=\frac{\frac{1}{2}\left(\bar{d}^2+\frac{n-k}{k-1}A_k\right)}{\frac{1}{2}(\bar{d}^2-A_k)}=\frac{\left(\bar{d}^2+\frac{n-k}{k-1}A_k\right)}{(\bar{d}^2-A_k)}.$$

On peut alors montrer que ([5]):

- Quand toutes les paires de points sont à égale distance, A_k vaut 0 et VRC vaut 1 (évident en regardant la définition de A_k).

- Le critère du $WGSS$ minimum maximise A_k pour un k donné.

- De plus, la fonction A_k peut aussi être utilisée pour comparer des partitions obtenues pour des nombres de classes différents:
  la différence $A_k - A_{k-1}$ indique un gain moyen dans la compacité intérieure des groupes résultant du passage de $k-1$ à k groupes.
  Donc, le comportement de A_k en fonction de k peut révéler l'existence de groupes.

Mettons cela en rapport avec l'indice VRC. Réecrivons:

$$\frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} = \frac{(1+\frac{n-k}{k-1}a_k)}{(1-a_k)} \qquad \text{où } a_k = \frac{A_k}{\bar{d}^2}.$$

On a alors que ([5]):

- a_k varie entre 0 et 1. a_k vaut 0 si toutes les paires de points sont à égale distance et 1 pour des groupements "idéaux" (où il n'y a pas de "variation" à l'intérieur des groupes[1].)

  En effet:

  - Il est évident que $a_k \geq 0$.
  - Si tous les points sont à égale distance, on a
    $A_k = 0 \quad \text{car } \bar{d}^2 = \bar{d}_i^2,\ i = 1, \ldots, k.$
    Ce qui entraine que: $a_k = 0$.
  - Si on a des groupements "idéaux": $\bar{d}_i^2 = 0,\ i = 1, \ldots, k.$

    $$\text{Or, } a_k = \frac{A_k}{\bar{d}^2} = \frac{\frac{1}{n-k}\left((n_1-1)(\bar{d}^2-\bar{d}_1^2)+\ldots+(n_k-1)(\bar{d}^2-\bar{d}_k^2)\right)}{\bar{d}^2}$$
    $$= \frac{\frac{1}{n-k}\bar{d}^2((n_1-1)+\ldots+(n_k-1))}{\bar{d}^2} + \frac{\frac{1}{n-k}\bar{d}^2\left((n_1-1)\bar{d}_1^2+\ldots+(n_k-1)\bar{d}_k^2\right)}{\bar{d}^2}$$
    $$= 1 + \frac{\frac{1}{n-k}\bar{d}^2\left((n_1-1)\bar{d}_1^2+\ldots+(n_k-1)\bar{d}_k^2\right)}{\bar{d}^2}.$$

    Et donc, a_k=1 .

- Si les points sont uniformément distribués dans l'espace, a_k va augmenter doucement et plus ou moins régulièrement avec k. VRC a quant à lui tendance a décroître quand k augmente si a_k est constant, ceci étant plus ou moins contrebalancé par l'augmentation de a_k.
  De toute façon, une distribution uniforme dans l'espace va généralement provoquer une variation régulière des valeurs de VRC.

1. Si tous les points sont différents, cela n'arrive que quand k atteint n.

- Par contre, si les points sont naturellement groupés en k_0 ensembles, le passage de $k_0 - 1$ à k_0 va provoquer une augmentation considérable de a_k et même de VRC (ce qui pourrait former une "bosse"). Plus précisément, le passage de $k_0 - 1$ à k_0 va beaucoup augmenter VRC si $\frac{a_{k_0}}{a_{k_0-1}}$ est supérieur au rapport $\frac{(k_0-1)}{(a_{k_0-1}+k_0-2)}$ (ce rapport n'est jamais plus petit que 1).

Nous savons maintenant que le calcul de VRC pour $k = 2, 3, \ldots$ aide à décider quel est le "meilleur" nombre de groupes. Calinski et Harabasz suggèrent de **choisir le nombre k pour lequel VRC a un maximum relatif ou absolu, ou au moins une croissance plus rapide.**
Si jamais les valeurs de VRC ont une croissance monotone avec celles de k, on peut conclure que la meilleure partition des points est celle où chaque point forme un groupe.

4.3.5 La méthode du "C-index" M_5 ([21])

Description

Définissons :

- $C(x_i, x_j) = \begin{cases} 1 \text{ si } x_i \text{ et } x_j \text{ sont dans la même classe.} \\ 0 \text{ sinon.} \end{cases}$

- $\Gamma = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}C(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}C(x_i, x_j)$

  où $d_{ij} \equiv$ carré de la distance euclidienne entre les objets i et j.

Remarquons que le "$\frac{1}{2}$" de la première égalité est intuitif. Il est là car on compte deux fois chaque distance entre deux points par symétrie.
C'est cette même symétrie qui permet de passer à la deuxième égalité.

Il est assez évident que :

$$\Gamma \equiv \sum_{i=1}^{k} \text{ des distances entre les objets du k-ième groupe.}$$

A partir de là, au moins cinq normalisations de Γ ont été suggérées comme indices pour déterminer le nombre de groupes. La normalisation utilisée ici est appelée C-index et est définie comme (Dalrymphe Alford, 1970) :

$$\text{C-index} = \frac{\Gamma - \min(\Gamma)}{\max(\Gamma) - \min(\Gamma)}$$

De plus, nous avons :

- $\min(\Gamma) \equiv \sum$des $(n_1! + n_2! + \ldots + n_k!)$ distances d_{ij} les plus petites .

- $\max(\Gamma) \equiv \sum$des $(n_1! + n_2! + \ldots + n_k!)$ distances d_{ij} les plus grandes .

En effet, pour k groupes de n_1, n_2, ..., n_k objets respectivement, il y a exactement $n_1! + n_2! + \ldots + n_k!$ distances entre des objets de même groupe ($n_1!$ pour le premier, ...). Donc, pour avoir les valeurs minimales (ou maximale) de Γ, il faut bien prendre les $n_1! + n_2! + \ldots + n_k!$ plus petites (ou plus grandes) distances.

Enfin, **la valeur minimale de l'indice est utilisée pour indiquer le nombre de classes** ([25]).

4.3.6 La méthode CCC M_6 ([29]; [31])

Le critère de classification cubique (C.C.C.) est le test statistique pour la détermination du nombre de classes utilisé dans le logiciel statistique SAS. L'indice est donné par la formule suivante :

$$CCC = \ln\left[\frac{1 - E(R^2)}{1 - R^2}\right] \frac{\sqrt{\frac{np}{2}}}{(0.001 + E(R^2))^{1.2}}$$

où

- R^2 est la proportion de variance expliquée par les classes.
- $E(R^2)$ est déterminé sous l'hypothèse que les données ont été générées par une disribution uniforme.
- n désigne le nombre d'observations.
- p est une estimation de la "dimensionnalité" déterminée par une analyse en composantes principales.

La valeur de l'indice correspondant au pic maximal indique le nombre de classes à retenir.
Notons aussi, que ces pics indiquent une bonne classification si la valeur de CCC ést plus grande que 2 ou 3. Tandis que des pics avec des valeurs de CCC entre 0 et 2 indiquent des classes possibles. Enfin, des valeurs très négatives (par exemple -30) peuvent être dûes à des points isolés.

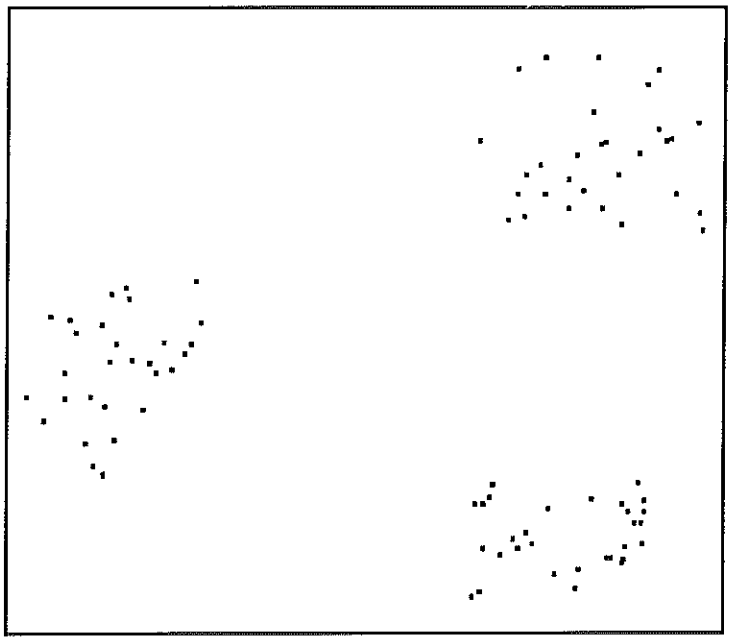
Enfin, ce critère n'est applicable que si on a des variables non-correlées. D'où l'utilité d'appliquer une analyse en composantes principales avant de l'utiliser.

4.4 Tableaux résultats et analyses

Pour commencer, remarquons que tous les jeux de données sont disponibles dans les annexes.
Attention : pour chaque exemple, à la fin des résultats, un petit commentaire sera fait sur les méthodes de détermination du nombre de classes basées sur les hypervolumes. Mais ce commentaire ne concernera que les résultats obtenus quand on applique ces méthodes de détermination du nombre de classes à la classification obtenue par la méthode des hypervolumes.

4.4.1 Données bien séparées

Cet exemple comporte 90 points. Rappelons qu'ils proviennent d'une simulation d'un processus de Poisson homogène dans trois classes bien séparées.



Trois classes bien séparées .

Données séparées (3 classes)		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	3	3	3	3	3	3
Voisin le plus éloigné	+	3	3	3	3	3	3
Moyenne	+	3	3	3	3	3	3
Ward	+	3	3	3	3	3	3

Tableau 1.

Il y a un "+" dans la deuxième colonne et cela pour toutes les méthodes de classification car elles fournissent les classes naturelles. Cela signifie que l'ensemble des points est fort bien structuré.
Dans les autres colonnes, on retrouve chaque fois le bon nombre de classes, c'est-à-dire 3. Donc, toutes les règles d'arrêt retrouvent le nombre exact de classes présentes, celles-ci étant bien les classes naturelles.

Dans cet exemple, nous remarquons tout de suite l'aspect pratique des méthodes M_1 et M_5: pour M_1, on doit chercher la valeur minimale de l'indice et pour M_5, la valeur maximale; mais nous savons que si les différents groupes sont bien marqués, cette valeur sera proche de 0 pour M_1 et de 1 pour M_5. C'est la cas ici (voir tableau 2). Tandis que pour M_4 et M_6, on doit simplement chercher la valeur maximale.

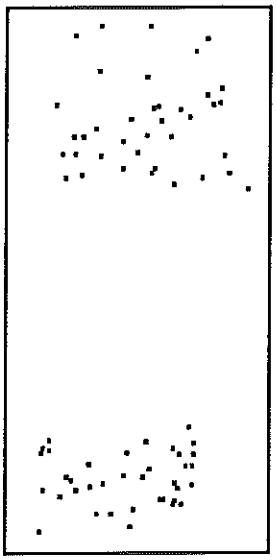
Méthode de ward	M_1	M_2	M_3	M_4	M_5	M_6
k= 7	0.9535	1.1699	0.9900	457.00	0.0044	14.587
6	0.9430	0.4438	0.5857	471.69	0.0061	15.696
5	0.9334	0.5505	0.6114	447.71	0.0105	15.668
4	0.9603	1.3347	0.9569	482.46	0.0078	17.260
3	1.0000	2.4006	1.7823	579.14	0.0000	21.015
2	0.9057	5.4668	6.0676	146.77	0.0407	3.710
1			3.7748	1.6308		0.000

Tableau 2: valeur des indices pour k=1,2,...,7 pour la méthode de Ward.

Rappelons enfin que les trois méthodes basées sur le critère des hypervolumes donnaient, elles-aussi, les bons résultats.

4.4.2 Données bien séparées bis

Cet exemple comporte 75 points.



Deux classes bien séparées .

Le premier exemple concernait déjà des données bien séparées. Seulement, rappelons-nous que dans l'article de Milligan et Cooper, beaucoup de méthodes parmi les meilleures avaient le plus de problèmes quand le nombre de classes présentes était deux. Cet exemple-ci permet juste de vérifier si toutes les méthodes donnent les bons résultats pour **deux** classes **bien séparées**.

Données séparées (2 classes)		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	2	2	2	2	2	2
Voisin le plus éloigné	+	2	2	2	2	2	2
Moyenne	+	2	2	2	2	2	2
Ward	+	2	2	2	2	2	2

Tableau 3.

A nouveau, toutes les méthodes trouvent le bon nombre de classes, celles-ci étant bien les classes naturelles (voir tableau 4 pour les valeurs des indices).

Donc, on ne remarque pas de problème pour deux classes si les données sont bien séparées.

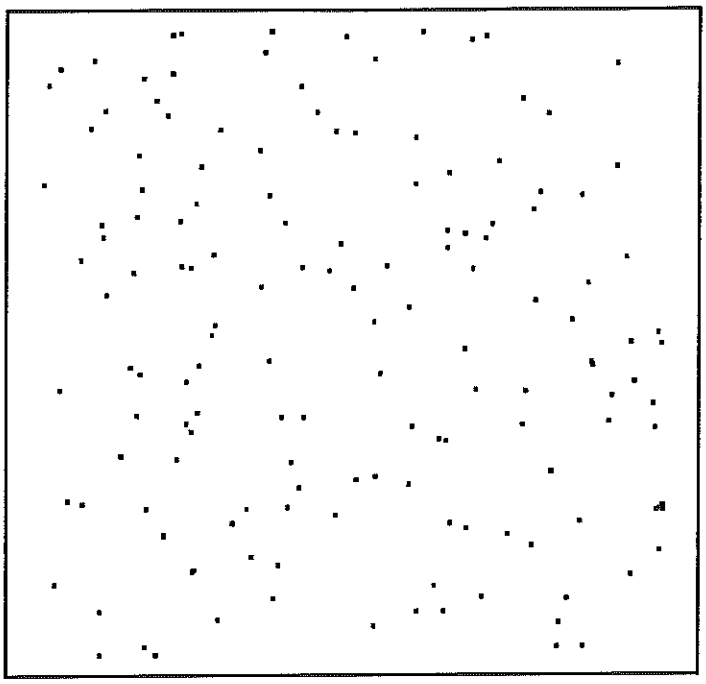
Méthode de la moyenne	M_1	M_2	M_3	M_4	M_5	M_6
k= 7	0.9405	0.9949	0.8980	357.49	0.0045	6.659
6	0.9169	1.6617	1.2653	304.73	0.0079	5.570
5	0.9125	1.9055	2.8867	322.48	0.0087	6.136
4	0.9098	2.5164	1.6624	262.76	0.0133	4.248
3	0.9204	0.2727	0.5179	340.39	0.0124	4.594
2	1.0000	0.2663	0.5137	530.27	0.0000	8.709
1		6.2964	7.0703			0.000

Tableau 4.

Les trois méthodes basées sur les hypervolumes donnent, ici aussi, les résultats corrects.

4.4.3 Données sans structure

Cet exemple comporte 150 données. Elles proviennent d'une simulation d'un processus de Poisson homogène dans un domaine du plan.



Données sans structure .

Données sans structure	M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	1	1	1	4 (1)	1	1
Voisin le plus éloigné	1	1 ou 3	1	4	1	1
Moyenne	1	1	1	4	1	1
Ward	1	1 ou 3	1	4	1	1

Tableau 5.

Examinons le tableau 5.
Pour commencer, la colonne qui indique si les classifications obtenues sont les classifications naturelles a bien évidement été supprimée car la classification en une classe ne peut être que celle contenant tous les points (et c'est bien la structure naturelle).
Examinons ensuite les autres colonnes:

- la méthode gamma, l'indice de Beale, la méthode CCC et le C-index (M_1, M_3, M_5 et M_6) retrouvent bien que les points ne forment qu'une

seule classe. Cela peut se voir dans le tableau 8 où M_1 n'a pas de maximum proche de 1, M_3 n'atteint jamais la valeur critique 5.30, M_5 n'a pas de minimum proche de 0 et M_6 n'a pas de maximum avant k=1.

- par contre, le critère de Duda-Hart (M_2) connait quelques petits problèmes.
  En effet, pour la méthode du voisin le plus éloigné et pour celle de Ward, la valeur de α utilisée par Milligan et Cooper dans leur article (3.20) ne suffit pas pour retrouver qu'il n'y a qu'une seule classe. Par contre, si α vaut 4 comme dans le travail de Gordon, nous retrouvons les résultats corrects. On voit effectivement dans le tableau 6 que la valeur trouvée pour M_2 est de 3.82. Ce qui correspond à une acceptation de l'hypothèse correspondant à la fusion des deux groupes concernés pour α valant 4, mais pas pour α valant 3.20.

- de son côté, la méthode de Calinski-Harabasz (M_4) semble incapable de retrouver l'absence de structure dans les points. Remarquons quand même que le résultat est un peu moins clair[1] quand on utilise la méthode du voisin le plus proche pour classifier les points. En effet, on voit dans le tableau 7 que le maximum pour M_4 en k valant 4 est plus marqué pour, par exemple, la méthode de la moyenne que pour la méthode du voisin le plus proche.

M_2	Voisin le plus éloigné	Méthode de Ward
k= 7	1.9306	1.9306
6	0.1666	0.1666
5	0.9497	0.9497
4	1.2033	1.2033
3	2.0182	2.0182
2	3.8214	3.8214
1	0.6669	0.6669

Tableau 6 : valeurs de M_2.

1. 1 est aussi un maximum local pour la valeur de l'indice mais il est plus petit.

M_4	Voisin le plus proche	Méthode de la moyenne
k= 7	10.602	104.04
6	7.8930	103.07
5	4.4089	96.951
4	5.1809	128.29
3	1.7898	111.23
2	2.2924	80.391
1		

Tableau 7 : valeurs de M_4.

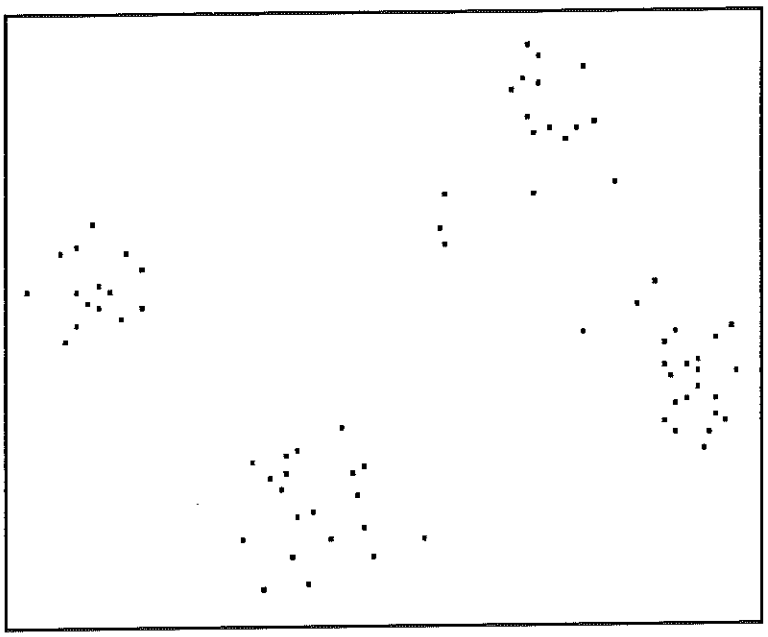
Voisin le plus proche	M_1	M_3	M_5	M_6
k= 7	0.3435	0.0293	0.3631	-24.196
6	0.2647	0.1331	0.4430	-24.368
5	0.3532	0.1341	0.4571	-24.478
4	0.3669	0.0135	0.4635	-22.623
3	0.2457	0.0790	0.5999	-21.623
2	0.3846	0.0086	0.5641	-13.026
1		0.0152		0.0000

Tableau 8.

Rappelons pour terminer qu'à nouveau, les méthodes basées sur les hypervolumes donnaient les bons résultats.

4.4.4 Données de Ruspini

Cet exemple comporte 75 données.



Données de Ruspini .

Données de Ruspini		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	4	4	(3 ou) 4	4	4	4
Voisin le plus éloigné	*	4	4	(3 ou) 4	4	4	4
Moyenne	+	4	4	(3 ou) 4	4	4	4
Ward	+	4	4	(3 ou) 4	4	4	4

Tableau 9.

Tout d'abord, remarquons qu'il y a un signe "*" dans la deuxième colonne pour la méthode du voisin le plus éloigné. Cela est du au fait que la classification obtenue est différente de celle obtenue par les autres méthodes mais reste tout à fait acceptable.

En ce qui concerne les résultats, toutes les méthodes retrouvent bien quatre classes à une exception près. En effet, la valeur de M_3 pour k=3 est de 4.65 (voir tableau 10). Or, cette valeur n'est suffisante pour dire qu'il y a quatre groupes que si on se limite à un niveau de précision de 0.01 (voir l'explication de M_3.

Enfin, deux petites remarques:

- pour M_1, la valeur de l'indice est un peu moins proche de 1 que pour les exemples des classes bien séparées. Par exemple, pour la méthode du voisin le plus proche, elle est ici de 0.9988 alors que pour deux classes bien séparées, elle était de 1.00. De même, pour M_5, la valeur de l'indice est un peu moins proche de 0 que pour le cas des classes bien séparées. A nouveau, pour la méthode du voisin le plus proche, elle est ici de 0.0009 alors que pour deux classes bien séparées, elle était de 0.0000. Nous tirerons parti de ces remarques plus tard dans les conclusions.

- pour les tests statistiques (M_2 et M_3 avec un niveau de précision de 0.01), nous avons procédé comme dans l'article de Milligan et Cooper en prenant la première valeur qui rejetait l'hypothèse correspondant à la fusion des deux groupes concernés. En effet, la valeur suivante de l'indice (qui aurait indiqué que l'ensemble de points contenait trois classes) correspondait aussi à un rejet de cette hypothèse. Le tableau 10 indique effectivement que M_2 et M_3 valent respectivement 4.32 et 7.09 pour k valant 2, ce qui correspond pour les deux à une acceptation de l'hypothèse correspondant à la fusion des deux groupes concernés.

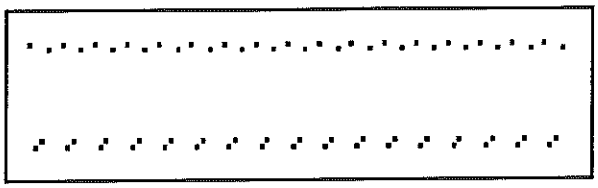
Méthode de la moyenne	M_2	M_3
k= 7	0.3707	0.5589
6	0.7098	0.6957
5	-0.8797	0.1799
4	1.4997	1.3170
3	4.2010	4.6568
2	4.3295	7.0933
1	3.5804	1.7090

Tableau 10.

Enfin, notons que les méthodes basées sur les hypervolumes retrouvaient à chaque fois quatre classes aussi.

4.4.5 Données avec deux classes parallèles

Cet exemple contient 68 données.



Deux classes parallèles .

Deux classes parallèles		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	1	1	1	6	1	1
Voisin le plus éloigné	-	1	2	1	2	1	1 ou 3
Moyenne	-	1	2	1	2	1	1
Wards	-	1	2	1	2	1	1 ou 3

Tableau 11.

Pour commencer, remarquons que seule la méthode du voisin le plus proche donne une bonne classification (c'était prévisible vu le biais des autres méthodes qui ont tendance à construire des classes hypersphériques).

Ensuite, nous voyons qu'aucune des méthodes ne retrouve le bon nombre de classes quand celles-ci sont les classes naturelles (voir tableau 12 pour des valeurs des indices).

Voisin le plus proche	M_1	M_2	M_3	M_4	M_5	M_6
k= 7	0.2913	2.0147	2.7405	10.816	0.2921	-17.685
6	0.3120	1.9526	1.8313	11.005	0.2879	-16.714
5	0.2935	1.5709	1.1935	9.6381	0.3079	-16.154
4	0.2279	1.1604	0.8673	6.3246	0.3716	-14.552
3	0.2011	-1.0414	0.2006	6.2634	0.4061	-8.7795
2	0.1685	-1.1475	0.1880	6.1740	0.4496	-9.8740
1		-2.4889	0.0907			0.0000

Tableau 12.

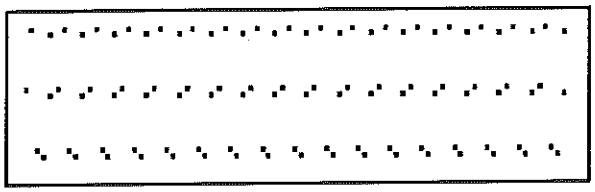
Maintenant, analysons M_2 et M_4 plus en détail car on remarque ici un biais de ces méthodes.

Pour commencer, quand la classification est bonne, elles ne retrouvent pas le bon nombre de classes. Par contre, quand la classification n'est pas bonne et que les classes trouvées ne sont pas les classes naturelles, M_2 et M_4 indiquent bien qu'il y a deux groupes. C'est un des problèmes dont nous avions parlé au chapite 2 et qui peut être très ennuyant. L'exemple suivant nous permettra de voir d'où vient ce problème et donc de décrire le "biais" de la méthode.

Enfin, notons que contrairement à toutes ces méthodes, la méthode des hypervolumes retrouve les bonnes classes et les méthodes de détermination du nombre de classes basées sur le critère des hypervolumes retrouvent le bon nombre de classes.

4.4.6 Données avec trois classes parallèles

Cet exemple contient 102 données.



Trois classes parallèles .

Trois classes parallèles		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	1	1	1	2	1	1
Voisin le plus éloigné	-	1	2	1	2	1	1
Moyenne	-	1	2	1	2	1	1
Ward	-	1	2	1	2	1	1

Tableau 13.

Comme dans l'exemple précédent, seule la méthode du voisin le plus proche fournit la bonne classification.

De plus, M_1, M_3, M_5 et M_6 ne parviennent toujours pas à donner des résultats corrects.

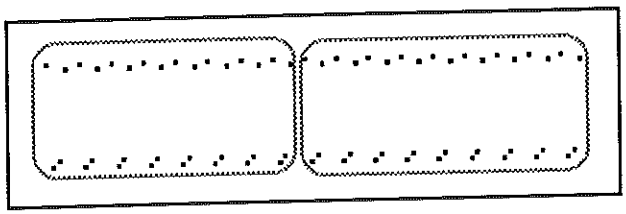
Ensuite, nous allons expliquer le biais dont nous parlions à l'exemple précédent. Effectivement, nous remarquons ici des résultats fort semblables à cet exemple. En fait, nous avions que :

- quand la classification était bonne, M_2 et M_4 indiquaient des mauvais résultats.
- quand la classification n'était pas bonne, M_2 et M_4 indiquaient qu'il y avait deux groupes mais ce n'étaient pas les classes naturelles.

Ici, la seule différence est que, quand la classification n'est pas bonne, M_2 et M_4 indiquent deux classes alors qu'il y en a trois.

A vrai dire, et cela pourrait alors expliquer tous ces résultats, on peut penser que M_2 et M_4 ont un problème que nous allons décrire de la manière suivante:

- Quand on a deux ou trois classes parallèles, les méthodes de classification qui ne retrouvent pas les classes naturelles donnent, pour k=2 (nombre de classes), des résultats tels que:



Les classes de départ étant assez proches, ces deux groupes sont assez compacts. Nous sommes donc faces à deux groupes de points hypersphériques et que l'on peut considérer comme formant deux classes. Or, M_2 et M_4 indiquent qu'il y a en effet deux groupes.

Donc, M_2 et M_4 **semblent privilégier les classes hypersphériques.**

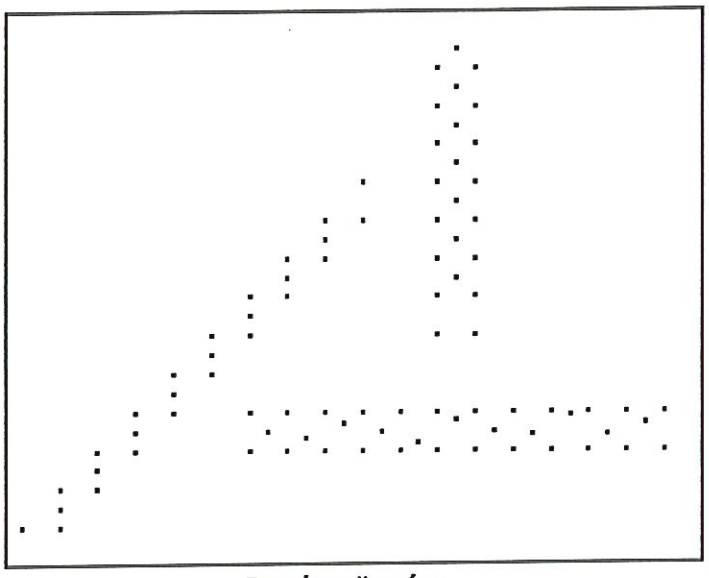
- Cela est "confirmé" par le fait que pour la méthode du voisin plus proche où les classes obtenues sont les bonnes mais sont allongées, M_2 et M_4 ne trouvent pas deux comme nombre de groupes.

N.B.: La notion de "privilégier les classes hypersphériques" n'est plus la même que pour les méthodes de classification.

Rappelons que la méthode et les critères basés sur les hypervolumes retrouvent à nouveau les bons résultats.

4.4.7 Données allongées

Cet exemple comporte 85 données.
On y remarque trois classes naturelles telles qu'aucune d'entre elles n'est séparable des autres par un hyperplan.



Données allongées .

Classes allongées		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	1	1	1	3	1	1
Voisin le plus éloigné	-	1	1 ou 3	1	1	1	1
Moyenne	-	1	1	1	1	1	1
Wards	-	1	1 ou 3	1	1	1	1

Tableau 14.

Voici un autre exemple où les classes sont allongées (mais plus parallèles). Il va confirmer ce que nous avons dit pour les deux exemples précédents.

Evidement, il n'y a encore que la méthode du voisin le plus proche qui retrouve la bonne classification.

Remarquons qu'à nouveau, le 3 dans la quatrième colonne est le nombre de classes trouvées avec α valant 3.20. Tandis que le 1 est la valeur trouvée pour α valant 4. Cela apparait clairement dans le tableau 15 où l'on voit que la valeur de M_2 pour k valant 2 est de 3.35 (ce qui est bien plus grand que 3.20 mais plus petit que 4.00).

M_2	Voisin le plus éloigné	Méthode de Ward
k= 7	2.1383	2.1383
6	2.1657	2.1657
5	2.4977	2.4977
4	2.3091	2.3091
3	1.7112	1.7112
2	3.3558	3.3558
1	1.7198	1.7198

Tableau 15 : valeurs de M_2.

Maintenant, passons à l'analyse du tableau 14. Tout d'abord, quasiment aucune méthode ne trouve les résultats corrects. En effet, quand M_2 retrouve bien trois classes, ce ne sont pas les classes naturelles. De plus, malgré que M_4 soit le seul qui trouve bien trois classes avec la méthode du voisin le plus proche, cela peut être dû au hasard. Effectivement, pour les deux exemples précédents (où les classes étaient allongées aussi), quand nous appliquions M_4 sur la classification obtenue par la méthode du voisin le plus proche, nous n'obtenions jamais un non plus comme nombre de classes mais nous obtenions un peu n'importe quoi: une fois deux alors qu'il y avait trois classes, une fois six alors qu'il y avait deux classes. Cela sera même confirmé à l'exemple suivant où nous trouverons cinq classes alors qu'il n'y en a que deux.
On peut donc maintenant affirmer qu'**aucune des six méthodes n'est capable de retrouver le bon nombre de classes si celles-ci sont allongées.**

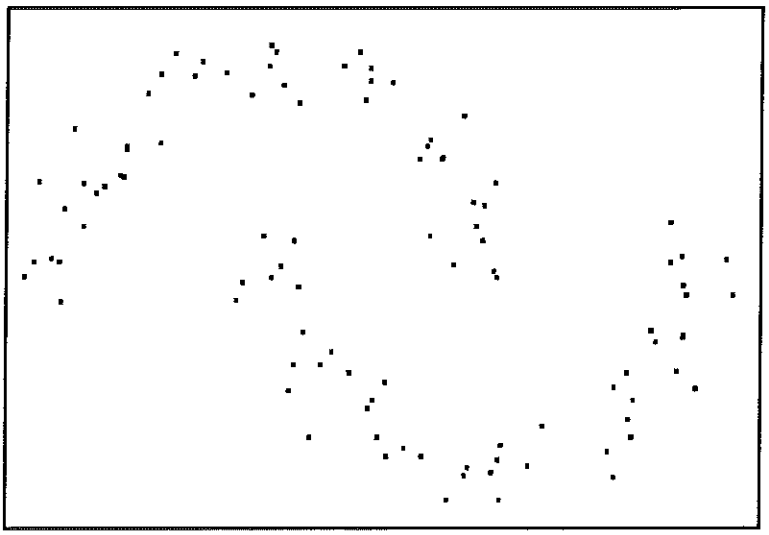
De plus, les classes hypersphériques trouvées par les trois méthodes de classification qui ne donnent pas les classes naturelles ne semblent pas assez compactes pour que M_2 et M_4 indiquent qu'il y a 2 ou 3 classes. En effet, cette fois, elles donnent un comme nombre de classes (comme les autres critères). Ce sera confirmé à l'exemple suivant.

Remarque: Pour tous les exemples où les classes sont allongées, les indices donnent quasiment toujours 1 comme nombre de classes (sauf M_2 et M_4 si les classes sont parallèles).

Notons que pour cet exemple, seules deux des trois méthodes basées sur les hypervolumes retrouvaient les bons résultats. La dernière ne donnant pas des résultats assez clairs pour pouvoir conclure.

4.4.8 Données en sourire

Cet exemple contient 100 données.
On peut remarquer que les classes naturelles ne sont pas convexes. De plus, elles ne sont pas séparables l'une de l'autre par un hyperplan.



Données en sourire .

Données en sourire		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	1	1	1	5	1	1
Voisin le plus éloigné	-	1	1	1	1	1	1
Moyenne	-	1	1	1	1	1	1
Wards	-	1	1	1	1	1	1

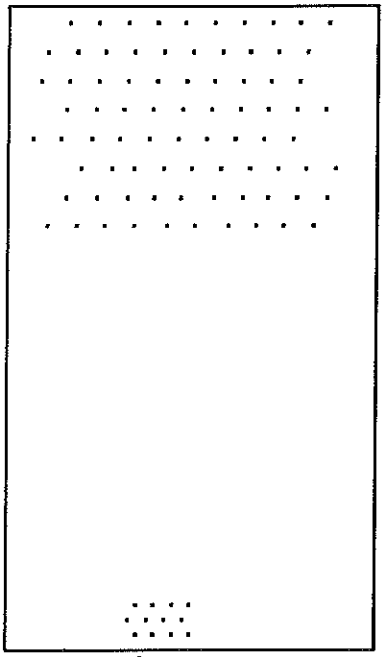
Tableau 16.

Les conclusions sont les mêmes que celles de l'exemple précédent.

Rappelons que les méthodes basées sur les hypervolumes ne parviennent pas non plus à retrouver les bons résultats. Les classes n'étant ni convexes, ni séparables par un hyperplan.

4.4.9 Données "gros-petit"

Cet exemple comporte 92 données. Il est composé de deux classes bien distinctes : une avec beaucoup de points et une avec peu de points.



Données "gros-petit".

Données gros-petit		M_1	M_2	M_3	M_4	M_5	M_6
Voisin le plus proche	+	2	2	1	2	2	2
Voisin le plus éloigné	+	2	2	1	2	2	3
Moyenne	+	2	2	1	2	2	6
Wards	+	2	2	1	2	2	3

Tableau 17.

Ce dernier exemple est assez intéressant. En effet, malgré que toutes les méthodes de classification donnent les bonnes classifications et que les classes sont bien séparées, les méthodes M_3 et M_6 ne parviennent pas retrouver qu'il y a deux classes (sauf pour M_6 appliqué à la classification obtenue par la méthode du voisin le plus proche). On peut se demander si cela est dû au fait qu'une classe comporte plus de points que l'autre ou alors si c'est à cause de l'espace occupé par la classe comportant le plus de points, qui est beaucoup plus grand que pour l'autre classe (cette hypothèse semble beaucoup moins probable).

Voici maintenant un exemple de valeurs de ces indices pour la méthode du voisin le plus éloigné.

Voisin le plus éloigné	M_3	M_6
k= 7	1.8049	5.4423
6	0.0368	5.3937
5	0.7210	5.1259
4	0.6612	5.5546
3	0.9076	5.8152
2	0.5146	3.7089
1	2.6617	0.0000

Tableau 18.

On remarque bien que :

- Pour M_3, la valeur de 5.30 n'est jamais dépassée et on obtient donc une classe. Notons quand-même que M_3 a un maximum pour k valant 1 mais pas assez grand pour dire qu'il y a deux classes.
- Pour M_6, la valeur maximale se situe en k valant 3.

Signalons aussi que M_1 et M_5 trouvent toujours exactement 1.0000 et 0.0000 comme valeurs pour k valant 2 (ces valeurs correspondent à une structure très nette).

Enfin, les méthodes basées sur les hypervolumes retrouvent quant à elles les résultats attendus.

4.5 Conclusions

La première chose à remarquer est qu'**aucune des 6 méthodes de l'article de Milligan et Cooper n'est capable de détecter les classes quand celles-ci sont allongées.** Et cela quelque soit la disposition (parallèles, non séparables par un hyperplan, ...) et le nombre (deux ou trois) de ces classes. De plus, quand la classification sur laquelle ces méthodes sont appliquées est bonne, cela montre que toutes ces méthodes privilégient les classes hypersphériques[1]. Notons encore que pour les classes allongées, M_1, M_3, M_5 et M_6 indiquent à chaque fois qu'il y a une seule classe.
Par contre, les méthodes basées sur le critère des hypervolumes sont quant à elles capables de détecter ces classes allongées, et ont à chaque fois donné les résultats attendus.

En ce qui concerne l'absence de structure, seule M_4 (et M_2 si α vaut 3.20) semble incapable de la détecter. Toutes les autres méthodes indiquent bien qu'il n'y a qu'une classe. Notons qu'en pratique, si une de ces méthodes (excepté celles basées sur le critère des hypervolumes) indique une classe, il faut essayer de vérifier que les classes n'étaient pas parallèles[2] (car dans ce cas, ces méthodes indiquent une classe également).

Nous avons aussi retrouvé le problème dont nous avions parlé au chapitre deux et que A. Hardy avait mentionné dans son travail: **certaines méthodes (M_2 si α vaut 3.20 et M_4) donnent parfois le bon nombre de classes alors que celles-ci[3] ne sont pas les classes naturelles.** Remarquons que cela est toujours arrivé **quand les classes étaient allongées.** Comme l'a signalé A.Hardy, il est alors nécessaire de **tenir compte des biais des méthodes de classification** pour essayer:

- De choisir des méthodes de classification adaptées aux données si c'est possible. Cela peut se faire de façon visuelle pour une dimension inférieure ou égale à trois. Par contre, cela demande un peu plus d'analyse pour des dimensions supérieures.

- De savoir si les classifications obtenues sont correctes, ou du moins quels biais elles présentent.

- D'appliquer plusieurs méthodes de classification qui n'ont pas les mêmes biais et d'analyser les résultats.

1. Rappelons-nous de la discussion concernant M_2 et M_4 lors de l'exemple 6.

2. C'est le seul cas que l'on connait où ces méthodes indiquent qu'il y a une classe quand ce n'est pas le cas mais il y en a peut-être des autres.

3. Les classes trouvées par la méthode de classification utilisée.

De plus, nous avons rencontré un autre problème : certaines méthodes (M_3 et M_6) semblent ne pas pouvoir retrouver le bon nombre de classes quand celles-ci ont des effectifs très différents. Par contre, M_1, M_2, M_4, M_5 et les méthodes basées sur le critère des hypervolumes ne rencontrent aucune difficulté dans ce cas. Rappelons notre interrogation concernant l'origine de ce problème. Est-ce dû au fait qu'une classe comporte plus de points que l'autre ou alors est-ce à cause de l'espace occupé par la classe comportant le plus de points, qui est beaucoup plus grand que pour l'autre classe (cette hypothèse semble beaucoup moins probable).

Maintenant, nous allons passer en revue toutes les méthodes. Nous pouvons néanmoins déjà souligner que **la méthode et les critères basés sur les hypervolumes sont les plus performants**. Les autres méthodes, quant à elles donnent des résultats assez satisfaisants (sauf peut être M_4).
Les méthodes Gamma, du C-index (M_1 et M_5) retrouvent des résultats tout à fait semblables. Elles ne donnent pas les bons nombres de classes pour des classes allongées. De plus, dans ce cas, elles indiquent qu'il n'y a qu'une seule classe.
Les méthodes de Beale et CCC (M_3 et M_6) retrouvent les mêmes résultats. Elles ont le même problème que M_1 et M_5 pour les classes allongées, c'est-à-dire qu'elles ne trouvent pas la bonne structure et qu'elles indiquent qu'il n'y a qu'une classe. De plus, elles ne retrouvent pas non plus le bon nombre de classes pour des classes d'effectifs très différents.
La méthode de Duda et Hart (M_2) ne retrouvent quant à elles pas de bons résultats si les classes sont allongées. Si α vaut 4, elle indique aussi qu'il y une seule classe présente. En outre, pour cette même valeur de α, elle peut retrouver l'absence de structure. Par contre, si α vaut 3.20, elle est incapable de retrouver l'absence de structure. De plus, pour cette valeur et pour des classes allongées, elle n'indique pas toujours qu'il n'y a qu'une classe et retrouve donc parfois des classes qui ne sont pas les classes naturelles, tout en privilégiant les classes hypersphériques.
La méthode de Calinski et Harabasz (M_4) semble être celle qui connait le plus de problèmes. Elle ne retrouve pas le bon nombre de classes quand celles-ci sont allongées. De plus, dans ce cas, elle n'indique pas un comme nombre de classes. Elle donne donc parfois le bon nombre de classes quand celles-ci ne sont pas les classes naturelles, tout en privilégiant les classes hypersphériques. Ensuite, elle ne retrouve pas l'absence de structure et ne donne pas de bons résultats dans le cas de classes d'effectifs très différents. Enfin, quand pour des classes allongées, la classification était bonne (avec la méthode du voisin le plus proche), elle semblait indiquer "n'importe quoi" comme nombre de classes.

Enfin, la méthode et les critères basés sur les hypervolumes retrouvent quant à eux chaque fois les bons résultats sauf dans le cas des données en sourire, où l'hypothèse de convexité des classes n'est pas respectée et où les classes ne sont pas séparables par un hyperplan.

Passons alors à des remarques un peu plus générales.
Tout d'abord, nous pouvons signaler l'aspect pratique de certaines méthodes :

- M_1 et M_5 cherchent respectivement une valeur maximale ou minimale, mais on sait que **pour une bonne classification, ces valeurs sont respectivement proches de 1 ou 0**. Ce qui n'est pas le cas de M_4 et M_6 qui cherchent simplement des valeurs maximales. Notons que pour M_6, certains "repères" existent ([31]).

- M_2 et M_3 cherchent une valeur bien précise.

Il serait peut-être intéressant de voir pour M_4 et M_6 si certaines valeurs correspondent à des classes bien structurées.
Ensuite, revenons sur la structure en classes allongées. Aucune des six méthodes n'est capables de la détecter. Mais si on regarde les indices d'un peu plus près, cela permet parfois de comprendre pourquoi. Par exemple, l'indice correspondant à M_1 compte "en gros" les points qui sont dans une même classe et qui sont plus éloignés que des points appartenant à des classes différentes. Or, pour des classes allongées, il est évident que ce nombre sera élevé. Peut-être serait-il intéressant de modifier l'indice pour n'utiliser que les k plus proches voisins. En tout cas, une analyse plus théorique reste possible. Elle a d'ailleurs déjà été faite en partie pour certains indices comme le C.C.C..
Signalons aussi un autre exemple qui pourrait poser des problèmes. C'est celui où les classes seraient d'effectifs très différents tout comme dans le jeu de données "gros-petit", mais où les classes seraient beaucoup plus proches. Outre M_3 et M_6, qui ne donnaient déjà pas les bons résultats quand les classes étaient éloignées, il est fort probable que d'autres méthodes ne donneront pas non plus les bons résultats. Surtout si la classe qui comporte le plus de points est beaucoup plus "volumineuse" que l'autre. En effet, on revient alors dans un cas similaire à celui des classes allongées.

Enfin, remarquons que le cas où k vaut 2 n'a pas vraiment posé de problème particulier dans nos exemples (sauf peut-être pour les données gros-petit ?). Or, c'est normalement le cas donnant les plus mauvais résultats ([25]). Peut-être serait-il intéressant de l'analyser plus particulièrement.

Annexe A

Partitions et hiérarchies

A.1 Partitions

Une partition de $E = \{x_1, x_2, \ldots, x_n\}$ en k classes $C_1, C_2, \ldots, C_k$ est définie par :

- $C_i \neq \emptyset \; i = 1, 2, \ldots, k$
- $C_i \cap C_j = \emptyset \; i, j = 1, 2, \ldots, k; \; i \neq j$
- $\sum_{i=1}^{k} C_i = E$.

Exemple :
$E = \{a, b, c\}$
$P = \{C_1, C_2\}$ où $C_1 = \{a\}$, $C_2 = \{b, c\}$.
On notera alors ici $P = a/bc$.

A.2 Familles de partitions

Une famille F de partitions est un ensemble de partitions $P_1, P_2, \ldots, P_n$ indicées par le nombre de classes.

Exemple :
$F = \{P_1, P_2, P_3, P_4\}$ où $P_1 = abcd$, $P_2 = ac/bd$, $P_3 = ab/c/d$ et $P_4 = a/b/c/d$
.

A.3 Hiérarchies de parties

Une hiérarchie H de parties de $E = \{x_1, x_2, \ldots, x_n\}$ est un ensemble de sous-ensembles de E qui vérifie :

- $E \subset H$
- $\forall x_i \in E : \{x_i\} \in H$
- $\forall E_1, E_2 \in H : E_1 \cap E_2 = \emptyset$
  ou $E_1 \subset E_2$
  ou $E_2 \subset E_1$.

Exemple :
$H' = \{P_1, P_2, P_3\}$ où $P_1 = abcd$, $P_2 = a/bcd$, $P_3 = a/bc/d$ et $P_4 = a/b/c/d$.

Annexe B

Jeux de données

B.1 Données bien séparées

4.48	6.93	4.36	8.12	5.01	7.22	4.77	6.87	4.21	7.20
4.86	6.72	3.99	5.04	3.55	7.37	4.01	6.16	4.12	8.00
5.12	6.76	3.44	7.60	5.61	7.54	5.34	7.04	3.35	6.71
3.96	7.52	3.85	5.18	5.52	8.22	4.18	5.62	4.00	5.06
4.10	6.89	3.36	6.29	2.74	6.33	3.01	5.93	3.71	5.56
4.67	6.12	3.79	6.33	4.41	7.94	5.45	7.19	3.12	7.65
10.23	10.52	12.12	11.00	10.86	9.66	12.59	9.14	13.18	10.71
11.70	9.88	11.00	9.96	10.87	11.72	12.53	9.96	11.84	10.29
11.70	9.41	13.19	11.68	13.84	10.81	12.32	10.49	13.41	10.55
10.71	9.24	13.31	10.53	13.49	9.64	13.00	11.45	12.87	10.31
11.32	9.64	12.24	10.46	11.31	11.89	12.26	9.42	11.23	10.12
13.92	9.05	12.18	11.90	11.96	9.69	10.98	9.29	13.89	9.32
10.13	3.03	10.42	4.67	12.12	4.64	12.60	4.55	12.42	3.64
11.40	4.47	11.14	3.88	12.72	4.43	12.82	4.24	10.30	3.80
10.60	3.71	12.37	3.64	12.97	4.42	12.60	3.57	12.64	3.62
10.26	3.09	11.89	3.48	12.87	4.90	12.97	4.61	12.67	3.84
10.48	4.87	11.03	4.07	10.88	3.82	10.17	4.55	12.94	3.90
11.51	3.40	11.85	3.16	10.30	4.55	12.93	4.23	10.80	3.98

B.2 Données bien séparées bis

12.12	9.99	10.84	9.98	12.21	9.32	12.38	10.25	11.29	11.11
13.11	9.23	12.72	10.45	10.64	9.65	10.53	10.52	12.12	11.00
10.86	9.66	12.59	9.14	13.18	10.71	11.70	9.88	11.00	9.96
10.87	11.72	12.53	9.96	11.84	10.29	11.70	9.41	13.19	11.68
13.44	10.81	12.32	10.49	13.41	10.55	10.71	9.24	13.31	10.53
13.49	9.64	13.00	11.45	12.87	10.31	11.32	9.64	12.24	10.46
11.31	11.89	12.26	9.42	11.23	10.12	13.92	9.05	12.18	11.90
11.96	9.69	10.98	9.29	13.59	9.32	10.73	4.03	10.42	4.67
12.12	4.64	12.60	4.55	12.42	3.64	11.80	4.47	11.14	3.88
12.72	4.43	12.82	4.24	10.30	3.80	10.60	3.71	12.37	3.64
12.97	4.42	12.60	3.57	12.64	3.62	10.26	3.09	11.89	3.48
12.87	4.90	12.97	4.61	12.67	3.84	12.18	4.17	11.73	4.07
10.88	3.82	12.07	4.05	12.94	3.90	11.51	3.40	11.85	3.16
10.30	4.55	12.93	4.23	10.80	3.98	11.11	4.27	12.64	3.93
10.43	4.52	11.27	3.38	12.73	3.57	10.27	4.45	11.37	3.92

B.3 Données sans structure

0.00	6.04	4.26	0.37	5.43	3.906	6.64	0.42	3.36	3.06
3.33	4.971	6.76	0.73	3.32	7.28	2.909	5.88	5.706	7.92
1.97	5.78	5.21	5.205	2.18	5.13	1.90	2.87	3.89	7.927
7.923	4.13	7.89	2.912	7.58	4.00	2.21	4.23	7.52	5.09
6.62	0.12	6.94	5.88	7.99	1.86	2.80	4.72	3.29	2.15
4.29	2.29	1.22	6.41	4.27	7.64	4.43	4.99	1.27	5.971
0.72	0.02	0.21	7.50	1.44	7.09	1.20	3.08	3.99	4.69
4.72	4.44	7.986	1.913	2.01	3.71	1.00	2.55	1.901	4.97
1.98	3.11	0.22	3.41	7.57	1.04	0.73	0.56	2.02	6.26
2.79	6.47	5.21	5.43	2.900	3.77	4.80	6.04	7.61	3.51
5.51	7.89	6.81	4.29	0.82	4.61	3.52	6.95	6.40	5.93
7.08	3.70	5.70	5.34	1.59	6.92	4.35	3.60	1.72	2.52
7.05	3.74	2.92	7.99	5.04	0.900	6.18	2.955	5.45	1.64
5.67	0.75	6.911	1.71	0.64	7.61	3.11	5.53	6.21	3.37
1.55	1.53	1.95	1.09	4.80	6.62	7.90	1.86	1.75	5.56
5.57	3.40	4.71	2.19	3.03	1.16	3.07	3.06	6.51	6.94
1.84	2.962	7.85	3.22	7.947	1.35	5.24	1.69	7.29	2.988
7.935	1.89	2.97	0.73	4.75	2.926	5.11	2.76	1.25	3.60
5.87	6.32	4.26	4.26	1.55	1.55	0.75	5.52	5.19	2.75
6.55	2.36	0.31	1.985	1.30	0.12	6.29	1.40	3.83	5.27
7.38	7.57	5.53	4.94	1.76	7.971	0.60	6.76	1.84	3.50
1.13	3.68	3.78	1.79	7.01	4.754	1.66	7.942	2.84	7.71
1.33	1.88	0.15	0.92	4.82	0.57	5.43	5.40	0.80	6.98
0.48	5.07	5.78	5.52	0.51	1.95	0.78	5.36	2.68	1.26
3.78	6.70	1.16	4.90	4.87	7.976	2.17	4.10	1.45	0.01
2.44	1.69	4.03	2.25	1.20	5.61	7.39	6.26	3.15	1.90
2.62	1.88	1.924	1.08	5.16	0.56	7.33	3.32	2.28	6.73
3.19	2.47	6.31	5.70	4.01	6.69	1.29	7.39	1.77	4.985
6.33	4.53	6.95	0.12	5.99	1.56	5.22	6.18	0.07	7.31
3.68	4.93	1.66	7.45	6.17	7.13	2.24	0.47	7.956	3.98

B.4 Données de Ruspini

45	17	50	23	41	26	74	26	51	30
62	34	63	39	46	37	43	40	51	42
125	42	127	48	127	51	122	51	120	45
131	56	119	55	122	57	118	61	127	62
103	63	116	72	94	88	77	82	100	98
102	100	105	101	90	107	95	108	95	113
2	71	9	62	11	71	13	69	15	68
17	71	20	78	23	68	53	18	65	23
57	26	63	28	54	31	61	38	48	35
49	38	49	41	59	46	126	45	129	47
124	53	120	50	118	47	124	56	118	57
124	58	120	63	130	64	113	68	109	90
78	88	78	79	94	99	97	100	93	102
92	109	103	111	93	115	8	78	11	65
11	79	14	83	15	72	19	66	23	75

B.5 Données avec deux classes parallèles

1.57333	1.66	1.62667	1.66	1.68000	1.66	1.73333	1.66	1.78667	1.66
1.84000	1.66	1.89333	1.66	1.94667	1.66	2.00000	1.66	2.05333	1.66
2.10667	1.66	2.16000	1.66	2.21333	1.66	2.26667	1.66	2.32000	1.66
2.37333	1.66	2.42667	1.66	1.55333	1.50	1.60667	1.50	1.66000	1.50
1.71333	1.50	1.76667	1.50	1.82000	1.50	1.87333	1.50	1.92667	1.50
1.98000	1.50	2.03333	1.50	2.08667	1.50	2.14000	1.50	2.19333	1.50
2.24667	1.50	2.30000	1.50	2.35333	1.50	2.40667	1.50	1.56333	1.51
1.61667	1.51	1.67000	1.51	1.72333	1.51	1.77667	1.51	1.83000	1.51
1.88333	1.51	1.93667	1.51	1.99000	1.51	2.04333	1.51	2.09667	1.51
2.15000	1.51	2.20333	1.51	2.25667	1.51	2.31000	1.51	2.36333	1.51
2.41667	1.51	1.54333	1.67	1.59667	1.67	1.65000	1.67	1.70333	1.67
1.75667	1.67	1.81000	1.67	1.86333	1.67	1.91667	1.67	1.97000	1.67
2.02333	1.67	2.07667	1.67	2.13000	1.67	2.18333	1.67	2.23667	1.67
2.29000	1.67	2.34333	1.67	2.39667	1.67				

B.6 Données avec trois classes parallèles

1.57333	1.66	1.62667	1.66	1.68000	1.66	1.73333	1.66	1.78667	1.66
1.84000	1.66	1.89333	1.66	1.94667	1.66	2.00000	1.66	2.05333	1.66
2.10667	1.66	2.16000	1.66	2.21333	1.66	2.26667	1.66	2.32000	1.66
2.37333	1.66	2.42667	1.66	1.57333	1.56	1.62667	1.56	1.68000	1.56
1.73333	1.56	1.78667	1.56	1.84000	1.56	1.89333	1.56	1.94667	1.56
2.00000	1.56	2.05333	1.56	2.10667	1.56	2.16000	1.56	2.21333	1.56
2.26667	1.56	2.32000	1.56	2.37333	1.56	2.42667	1.56	1.55333	1.47
1.60667	1.47	1.66000	1.47	1.71333	1.47	1.76667	1.47	1.82000	1.47
1.87333	1.47	1.92667	1.47	1.98000	1.47	2.03333	1.47	2.08667	1.47
2.14000	1.47	2.19333	1.47	2.24667	1.47	2.30000	1.47	2.35333	1.47
2.40667	1.47	1.56333	1.46	1.61667	1.46	1.67000	1.46	1.72333	1.46
1.77667	1.46	1.83000	1.46	1.88333	1.46	1.93667	1.46	1.99000	1.46
2.04333	1.46	2.09667	1.46	2.15000	1.46	2.20333	1.46	2.25667	1.46
2.31000	1.46	2.36333	1.46	2.41667	1.46	1.53333	1.57	1.58667	1.57
1.64000	1.57	1.69333	1.57	1.74667	1.57	1.80000	1.57	1.85333	1.57
1.90667	1.57	1.96000	1.57	2.01333	1.57	2.06667	1.57	2.12000	1.57
2.17333	1.57	2.22667	1.57	2.28000	1.57	2.33333	1.57	2.38667	1.57
1.54333	1.67	1.59667	1.67	1.65000	1.67	1.70333	1.67	1.75667	1.67
1.81000	1.67	1.86333	1.67	1.91667	1.67	1.97000	1.67	2.02333	1.67
2.07667	1.67	2.13000	1.67	2.18333	1.67	2.23667	1.67	2.29000	1.67
2.34333	1.67	2.39667	1.67						

B.7 Données allongées

1.	1.	2.	1.	2.	2.	3.	2.	3.	3.
4.	3.	5.	4.	5.	5.	6.	5.	6.	6.
7.	3.	7.	4.	7.	6.	7.	7.	8.	3.
8.	4.	8.	7.	8.	8.	9.	3.	9.	4.
9.	8.	9.	9.	10.	3.	10.	4.	10.	9.
10.	10.	11.	3.	11.	4.	12.	3.	12.	4.
12.	6.	12.	7.	12.	8.	12.	9.	12.	10.
12.	11.	12.	12.	12.	13.	13.	3.	13.	4.
13.	6.	13.	7.	13.	8.	13.	9.	13.	10.
13.	11.	13.	12.	13.	13.	14.	3.	14.	4.
15.	3.	15.	4.	16.	3.	16.	4.	17.	3.
17.	4.	18.	3.	18.	4.	7.5	3.5	8.5	3.3
9.5	3.7	10.5	3.5	11.5	3.2	12.5	3.8	13.5	3.5
14.5	3.4	15.5	3.9	16.5	3.4	17.5	3.7	12.5	7.5
12.5	8.5	12.5	9.5	12.5	10.5	12.5	11.5	12.5	12.5
12.5	13.5	2.	1.5	3.	2.5	5.	4.5	6.	5.5
7.	6.5	8.	7.5	9.	8.5	4.	3.5	4.	4.

B.8 Données en sourire

3.03749	2.08972	3.01959	2.11360	2.84041	2.14835	2.94592	2.32273
2.97207	2.25784	2.97850	2.41149	3.03286	2.51179	2.73136	2.27549
2.72975	2.71330	2.78165	2.62101	2.92549	2.42649	2.88239	2.82270
2.78689	2.63322	2.67921	2.62151	2.72049	2.68161	2.46044	3.03574
2.55941	2.97249	2.43778	2.88818	2.33448	3.04500	2.45439	2.97990
2.40827	3.10743	2.13212	2.87629	2.02023	3.10982	2.05756	2.95561
1.99680	3.04692	2.00106	3.13467	1.91087	2.90987	1.79698	3.01289
1.65148	3.00094	1.68947	3.06231	1.56870	3.10504	1.50411	3.00967
1.44260	2.92068	1.50275	2.69405	1.34608	2.66919	1.34784	2.68142
1.11398	2.75796	1.31790	2.54864	1.33780	2.54085	1.15175	2.51401
0.95454	2.52368	1.24538	2.50266	1.06596	2.40284	1.20864	2.47407
1.15333	2.31807	1.04184	2.15729	0.92931	2.15855	1.01266	2.17444
0.88934	2.09182	1.04967	1.97989	1.97185	2.27974	2.11182	2.25781
2.05393	2.13898	1.88092	2.06591	1.85163	1.98431	2.13474	2.04519
2.01125	2.08862	2.10791	1.69614	2.15545	1.83981	2.28026	1.75617
2.08820	1.57978	2.23043	1.69960	2.36499	1.65885	2.44829	1.50030
2.52634	1.62005	2.18283	1.36992	2.47398	1.53694	2.49632	1.36780
2.53709	1.28101	2.61891	1.32204	2.80921	1.08642	2.69863	1.28269
3.05937	1.33163	2.90662	1.23645	2.89489	1.19962	3.04129	1.26645
3.01197	1.20938	3.17961	1.24294	3.05367	1.08800	3.24804	1.42186
3.63755	1.45096	3.54171	1.30403	3.57400	1.19066	3.65959	1.53642
3.65309	1.36909	3.57058	1.59867	3.63100	1.66048	3.75925	1.79951
3.85134	1.66683	3.94241	1.58877	3.88385	1.82265	3.88180	2.05372
3.73750	1.85008	4.10882	2.00944	3.88437	1.83018	3.82376	2.15118
3.87529	2.17916	3.89741	2.01047	4.08177	2.16492	3.82280	2.33583

B.9 Données "gros-petit"

1.5502	1.75	1.5703	1.75	1.5904	1.75	1.6105	1.75	1.6306	1.75
1.6507	1.75	1.6705	1.75	1.6904	1.75	1.7103	1.75	1.7305	1.75
1.5454	1.73	1.5653	1.73	1.5852	1.73	1.6056	1.73	1.6257	1.73
1.6456	1.73	1.6655	1.73	1.6854	1.73	1.7053	1.73	1.7253	1.73
1.5654	1.77	1.5856	1.77	1.6055	1.77	1.6254	1.77	1.6453	1.77
1.6654	1.77	1.6852	1.77	1.7054	1.77	1.7253	1.77	1.7452	1.77
1.5406	1.69	1.5600	1.69	1.5805	1.69	1.6004	1.69	1.6200	1.69
1.6403	1.69	1.6602	1.69	1.6800	1.69	1.7005	1.69	1.7203	1.69
1.5634	1.71	1.5832	1.71	1.6031	1.71	1.6232	1.71	1.6435	1.71
1.6637	1.71	1.6835	1.71	1.7034	1.71	1.7233	1.71	1.7432	1.71
1.5504	1.63	1.5703	1.63	1.5902	1.63	1.6104	1.63	1.6334	1.63
1.6526	1.63	1.6756	1.63	1.6945	1.63	1.7143	1.63	1.7354	1.63
1.5633	1.65	1.5844	1.65	1.6055	1.65	1.6243	1.65	1.6424	1.65
1.6656	1.65	1.6854	1.65	1.7034	1.65	1.7232	1.65	1.7445	1.65
1.5733	1.67	1.5944	1.67	1.6100	1.67	1.6300	1.67	1.6500	1.67
1.6700	1.67	1.6900	1.67	1.7100	1.67	1.7300	1.67	1.7500	1.67
1.6125	1.37	1.6250	1.37	1.6375	1.37	1.6500	1.37	1.6125	1.35
1.6250	1.35	1.6375	1.35	1.6500	1.35	1.6075	1.36	1.6200	1.36
1.6325	1.36	1.6450	1.36						

Bibliographie

[1] ANDERBERG, M.R. (1973). Cluster analysis for applications. Academic Press, New York.

[2] BAKER, F.B. et HUBERT, L.J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70, 31-38.

[3] BEALE, E.M.L. (1969). Cluster analysis. London: Scientific Control Systems.

[4] BOCK, H.H. (1985). On some significance tests in cluster analysis. *Journal of Classification, 2, 77-108.*

[5] CALINSKI, R.B. et HARABASZ, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1-27.

[6] CHANDON, J.L. et PINSON, S. (1981). Analyse typologique: théorie et applications. Masson, Paris.

[7] DALRYMPLE-ALFORD, E.C. (1970). The measurement of clustering in free recall. *Psychological Bulletin*, 75, 32-34.

[8] DIDAY, E. (1972). Nouvelles méthodes et nouveaux concepts en classification automatique. Thèse d'Etat, Paris.

[9] DELATTRE, M. (1979). Classification optimale bicritère: méthodes, algorithmes et applications. Thèse de doctorat, Faculté Universitaire de Mons, Belgique.

[10] DUDA, R.O. et HART P.E. (1973). Pattern classification and scene analysis. Wiley-Interscience, New York.

[11] EVERITT, B. (1980). Cluster analysis. Halsted press, London.

[12] FISHER, L. et VAN NESS, J.W. (1971). Admissible clustering procedures. *Biometrika*, 58.

[13] GORDON A.D. (1981). Classification. Chapman and Hall, London.

[14] GORDON, A.D. (1997) How many clusters? An investigation of five procedures for detecting nested cluster structure. *Actes de IFCS-96 Conference, Kobe, à paraître.*

[15] HARDY, A. et RASSON, J.P. (1982). Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des Données, 7, 41-56.*

[16] HARDY, A. (1983). Statistique et classification automatique. Un modèle - Un nouveau critère - Des algorithmes - Des applications. Thèse de doctorat, F.U.N.D.P., Namur, Belgique.

[17] HARDY, A. (1992). On tests concerning the existence of a classification. *Belgian Journal of Operations Research, Statistics and Computer Science, Vol 31, nr 3-4.*

[18] HARDY, A. (1996). On the number of clusters. *Computational Statistics and Data Analysis, 23, 83-96.*

[19] HODSON, F.R., SNEATH, P.M.A. et DORAN J.E. (1966). Some experiments in numerical analysis of archeological data. *Biometrika*, 53, 311-324.

[20] HUBERT, L.J. et BAKER, F.B. (1975). Measuring the power of hierarchical cluster anlysis. *Journal of the American Statistical Association*, 70, 31-38.

[21] HUBERT, L.J. et LEVIN, J.R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83, 1072-1080.

[22] JAMBU, M. (1972). Techiques de classification automatique appliquées à des données de sciences humaines. Thèse de Doctorat de 3ème Cycle, Paris.

[23] JARDINE, N. et SIBSON, R. (1971). Mathematical Taxonomy. J.W.Wiley, New York.

[24] LINNE, K. (1737). Genera Plantarum.

[25] MILLIGAN, G.W. et COOPER, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159-179.

[26] MOORE, M. (1984). On the estimation of a convex set. *The Annals of Statistics*, 12, 3, 1090-1099.

[27] NEVEU, J. (1974). Processus ponctuels. Technical Report, Laboratoire de Calcul des Probabilités, Université de Paris VI.

[28] RASSON, J.P. (1979). Estimation de formes convexes du plan. *Statistique et Analyse des Données, 1, 31-46.*

[29] RAY, A.A. (1982). SAS User's Guide: Statistics. Cary, North Carolina: SAS Institute.

[30] RIPLEY, B.D., et RASSON, J.P. (1977). Finding the edge of a Poisson Forest. *Journal of Applied Probability, 14, 483-491.*

[31] SARLE, W.S. (1983). Cubic Clustering Criterion. Technical Report: A-108, SAS Institute Inc..

[32] SIDAK, Z. (1979). Sone ideas for the comparison of clustering procedures. Technical report, Mathematical Institute, Czech. Acad. Sci., Prague.

[33] CLUSTAN analysis software (1987). Computing Laboratory, Université St Andrews, St Andrews, Ecosse.